Análisis de series temporales

Práctico de Análisis y Visualización

En este práctico se propone explorar un dataset que contiene información de varias agencias de correo de Brasil, con el objetivo de extraer toda la información relevante sobre estos datos y presentarla de manera organizada y sencilla.
En estos datos vamos a encontrar **información geográfica** de los compradores y vendedores, el tipo de **servicio** por el que viajan los paquetes, el **estado** del paquete, algunas **fechas relevantes** y la cantidad de días hábiles que tardó el envío en llegar a su destino (**target**).

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, explicada con títulos y texto en markdown en cada sección, como se muestra en el informe de ejemplo.
El informe debe tener al menos las siguientes secciones: Introducción,Información obligatoria, información extra y Conclusión.
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

Se espera que se presente al menos la siguiente información.

En general

- Cantidad y proporción de **envíos**, **servicios, tipos de envíos** y **rutas** (consideramos como ruta la tripla zipcode, zipcode, servicio).
- Puntos máximos y mínimos de cada feature.

- Calcular estadísticos como la media, mediana, desviación estándar y percentiles del **target**.
- Graficar la distribución del **target** ¿Responde a alguna distribución conocida?
- Graficar solo la parte más informativa de la distribución del **target**, teniendo cuidado con elegir correctamente los parámetros de los gráficos, como la cantidad de bins en un histograma.
- Identificar y graficar outliers. ¿Son significativos?
- ¿Los fines de semana son diferentes a los días de semana? ¿En qué sentido?
- ¿Existe algún periodo de tiempo diferente a los demás? Comparar gráficamente las distribuciones de los targets (Puede ayudar utilizar información externa).
- Observando la distribución del **target** semana a semana. Explicar que sucede y cuál puede ser la razón. Graficar las distribuciones en conjunto o la diferencia entre ellas.
- ¿Existen rutas más representadas que otras?
- La cantidad de items por paquete, ¿tiene relación con la velocidad del envío?
- ¿Existen variables correlacionadas?
- Graficar la distribución del target agrupando por tipo de envío.
- Determinar cuales son los servicios y los estados más representados

Servicios

- Graficar solo la parte más informativa de la distribución del **target** para los 4 servicios más representados, con los cuidados correspondientes.
- ¿Existen servicios más rápidos que otros?
- ¿Existen servicios similares entre sí? ¿Alguna idea de porque?
- (Extra) ¿Podrías identificar cuántos correos aparecen en los datos?

Estados

- Graficar solo la parte más informativa de la distribución del **target** para los 4 estados más representados, con los cuidados correspondientes.
- ¿Cómo están distribuidos los vendedores geográficamente?
- ¿Cómo es la participación de los servicios dentro y fuera de San Pablo?
- (Extra) ¿Podrías explicar porque estos estados están más representados que los otros?
- Realizar un mapa de calor utilizando los zipcodes de los vendedores y los compradores.
- ¿Cual es la relación entre los zipcodes y los estados?

Se espera que utilicen las preguntas como guía de lo que se espera que vean en los datos, todas las respuestas 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.**