Licenciatura en Ciencia de Datos
ECYT_UNSAM

Introducción a la Ciencia de Datos - 2C 2022

Guía de Trabajos Prácticos Nº5

Hoy vamos a trabajar con dos bases de datos que hablan de las condiciones meteorológicas durante la Segunda Guerra Mundial. Pueden descargarlas y ver la documentación acá.

La meta final es encontrar si hay relación entre las temperaturas mínimas y máximas diarias y eventualmente pensar cómo predecir la temperatura mínima sabiendo la máxima.

Notarán que les soltamos un poco las manos; con lo visto en la teórica y la bibliografía queremos que se lancen a buscar y probar alternativas para cumplir con los objetivos.

1. Familiaricense y **limpien** ambas bases de datos. ¿Qué información contiene cada una de ellas? ¿Hay datos faltantes? ¿se entienden las unidades de las variables? ¿Hay outliers?... Recuerden las pautas presentadas en las semanas anteriores.
2. Quédense con **una** estación meteorológica después de haber **unido** ambos datasets (piensen con cuidado el método de unión).
3. **Grafiquen** una temperatura en función de otra. ¿Puede observarse una **tendencia** clara? ¿A qué familia de modelos pertenecería?
4. Determinen si hay puntos que se **desvían** del comportamiento observado en el punto 3. Decidan qué hacer con ellos.
5. **Ajusten** un modelo a sus datos. ¿Qué parámetros tendrá este ajuste? ¿Cuáles son las variables? Reporten el valor de los parámetros obtenidos.
6. ¿Existe alguna otra **variable** en la base de datos que sea (o pueda convertirse en) **categórica** para considerar en nuestro modelado?
    a. Identifiquen la variable
    b. Apliquen otro modelo teniendola en cuenta
    c. Observen el ajuste resultante. 'Todos los modelos son incorrectos', pero ¿podemos determinar si alguno es 'mejor'?

7. Hacer un **análisis de los residuos** de ambos modelos. ¿Nos permite esto responder la pregunta 6 de una manera más eficiente?

8. Retomemos los datos de todas las estaciones (eliminemos el filtro del punto 2).
    a. Repetir los ajustes tomando la estación como variable categórica. *Para esto pensar si vamos a usar el código o el nombre y que implica cada uno para el ajuste.*
    b. Incluyan en el gráfico la recta identidad (y=x). ¿Qué les permite decir sobre las tendencias encontradas?
9. Busquen temperaturas mínimas actuales en la estación de máxima y mínima elevación. ¿Pueden predecir la temperatura máxima con sus ajustes?

TIP

Pueden armar una variable categórica a partir de una numérica usando la función `case_when`:

```
df %>% mutate(categorica = case_when(
     numerica > 5 ~ “mayor a 5”,
     numerica <= 5 ~ “menor o igual a 5”,
     TRUE ~ NA,
  ))
```