Segundo Parcial - INF-A - CUNEF Universidad

Análisis y Explotación de la Información

2023-11-27

Ejercicio 1 (3 pts)

Los datos `fish_encounters`, disponibles en tidyverse, contienen información de diversas estaciones para la monitorización de diferentes especies de peces en diferentes ríos. En particular, cada estación registró qué tipo de peces ha detectado. La base de datos contiene esta información:

- `fish`: identificador del tipo de pez
- `station`: identificador de la estación
- `seen`: = 1 si el pez fue detectado en la estación.

Así por ejemplo, la estación MAE ha detectado los peces 4842, 4843, 4844, 4858 y 4861; y para estos peces el valor de seen es 1.

Imaginemos que nuestra unidad observacional son cada uno de los diferentes tipos de pez y nuestras variables son el nombre del pez, y todas las estaciones. Ordena estos datos para que tengan esta estructura.

El dataframe resultante ha de tener por tanto una columna llamada `fish` que contiene el nombre del pez, una columna con el nombre de la primera estación, otra con el de la segunda, etc. Para la variable referente a la estación X, en la fila referente al pez Y, habrá un 1 si la estación X detectó al pez Y y un 0 en caso contrario.

Pista: Puedes utilizar: `df %>% replace(is.na(.), 0)`, este código reemplaza todos los NAs del dataframe `df` por 0.

Ejercicio 2 (3.5 pts)

La base de datos presente en el fichero `states.csv` contiene la siguiente información acerca de los estados de EEUU

- `state`: nombre completo del estado
- `abb`: siglas del estado
- `region`: región geográfica del estado
- `population`: población del estado

La base de datos del fichero `results_us_election_2016.csv` contiene la siguiente información sobre las elecciones de 2016 por estados:

- `state`: nombre completo del estado
- `electoral_votes`: número de votos del estado para la elección presidencial
- `clinton`: porcentaje de apoyo a Hillary Clinton
- `trump`: porcentaje de apoyo a Donald Trump
- `others`: porcentaje de apoyo a otras candidaturas.

Visualiza (con una gráfica) el porcentaje de apoyo a la candidatura de Trump según la región geográfica de los estados. Extrae dos conclusiones. ¿Qué estado del Sur tiene un comportamiento atípico?

Ejercicio 3 (3.5 pts)

Los datos presentes en el siguiente link contienen información acerca de la evolución del desempleo en EEUU desde 1948 hasta hoy. La base de datos tiene dos variables: `DATE` la fecha en formato año, mes día y `UNRATE`, la tasa de desempleo. Responde a las siguientes cuestiones.

```
link <- "https://fred.stlouisfed.org/graph/fredgraph.csv?bgcolor=%23e1e9f0&chart_type=line&drp=0&fo=ope
```

- Descarga los datos e impórtalos a R.
- Añade una nueva columna a la base de datos que se llame `mes` y contenga el mes de cada observación con etiquetas (es decir, Enero aparecerá como Jan (o Ene si tu R está en castellano), Febrero con Feb/, etc.)
- Utilizando únicamente los datos posteriores al año 2019, calcula el desempleo **medio** por mes. Crea un gráfico de barras cuyo eje x corresponda a la variable `mes` y el eje y contenga el desempleo medio. ¿Qué mes tiene mayor tasa de desempleo media? ¿Sabrías explicar por qué?

Pista: recuerda añadir `stat="identity"` para que en el eje y del gráfico de barras de la última parte aparezca el desempleo medio.