prog_datasci_6_preproc_entrega

May 6, 2020

1 Programación para *Data Science*

1.1 Unidad 6: Preprocesamiento de datos en Python

2 Introducción

En este Notebook encontraréis el conjunto de **actividades evaluables** como PEC de la asignatura.

Or recordamos que os miréis antes de abordar este notebook los ejercicios para practicar. Éstos no puntúan para la PEC, pero os recomendamos que los intentéis resolver como parte del proceso de aprendizaje. Encontraréis ejemplos de posibles soluciones a los ejercicios al propio notebook, pero es importante que intentéis resolverlos vosotros antes de consultar las soluciones. Las soluciones os permitirán validar vuestras respuestas, así como ver alternativas de resolución de las actividades. También os animamos a preguntar cualquier duda que surja sobre la resolución de los **ejercicios para practicar** en el foro del aula.

En relación a las actividades evaluables de este notebook, veréis que cada una de ellas tiene asociada una puntuación que indica el peso que tiene la actividad sobre la nota de la PEC. Adicionalmente, hay un ejercicio opcional, que no tiene puntuación dentro de la PEC, pero que se valora al final del semestre de cara a conceder las matrículas de honor y redondear las notas finales. Podéis sacar la máxima nota de la PEC sin necesidad de hacer este ejercicio! El objetivo de este ejercicio es que sirva como pequeño reto para los estudiantes que quieran profundizar en el contenido de la asignatura.

Además, veréis que todas las actividades tienen una etiqueta que indica los recursos necesarios para llevarla a cabo. Hay tres posibles etiquetas:

- NM **Sólo materiales**: las herramientas necesarias para realizar la actividad se pueden encontrar en los materiales de la asignatura.
- EG **Consulta externa guiada**: la actividad puede requerir hacer uso de herramientas que no se encuentran en los materiales de la asignatura, pero el enunciado contiene indicaciones de dónde o cómo encontrar la información adicional necesaria para resolver la actividad.
- EI **Consulta externa independente**: la actividad puede requerir hacer uso de herramientas que no se encuentran en los materiales de la asignatura, y el enunciado puede no incluir la descripción de dónde o cómo encontrar esta información adicional. Será necesario que el estudiante busque esta información utilizando los recursos que se han explicado en la asignatura.

Es importante notar que estas etiquetas no indican el nivel de dificultad del ejercicio, sino únicamente la necesidad de consulta de documentación externa para su resolución. Además, recordad que las **etiquetas son informativas**, pero podréis consultar referencias externas en cualquier momento (aunque no se indique explícitamente) o puede ser que podáis hacer una actividad sin consultar ningún tipo de documentación. Por ejemplo, para resolver una actividad que sólo requiera los materiales de la asignatura, puedéis consultar referencias externas si queréis, ya sea tanto para ayudaros en la resolución como para ampliar el conocimiento!

En cuanto a la consulta de documentación externa en la resolución de los ejercicios, recordad **citar siempre la bibliografía utilizada** para resolver cada actividad.

3 Ejercicios para la PEC

A continuación, los **ejercicios y preguntas teóricas que debe completar en esta PEC** y que forman parte de la evaluación de esta unidad.

4 Ejercicio 1

La comisión europea ha hecho un esfuerzo en recopilar los datos en relación a la pandémia del sars-cov-2. La principal organización encargada es la European Center for Disease Prevention and Control (ECDC, https://www.ecdc.europa.eu). Desde el comienzo de la pandemia de coronavirus, el equipo de epidemiólogs del ECDC ha recopilado diariamente el número de casos y muertes de COVID-19, según informes de las autoridades sanitarias de todo el mundo, datos con los que trataremos en este notebook.

Importa el fichero `covid19.csv` de la carpeta de datos como un objeto pandas. Usa el formato que mejor se adecue para las diferentes columnas. Una vez importados los datos, realiza las siguientes operaciones:

a) Muestra por pantalla el número total de medidas y el nombre de las variables. Muestra las 5 primeras entradas. Explica qué captura cada variable.

(1 punto) NM

[1]:
```
# Respuesta
```

b) Buscar en los valores de casos y muertes, valores claramente anómalos o sin sentido. En caso de encontrar valores anómalos sustitúye estos por NaN.

(1.5 puntos) EI

[4]:
```
# Respuesta
```

c) Elimina las muestras de las que no tengamos información sobre la población (popData2018).

(0.5 puntos) NM

[10]:
```
# Respuesta
```

d) Observa la codificación de las variables `geoId`, `countryterritoryCode` y `continentExp`. Corrige la codificación según tu mejor criterio si observas anomalías, razona y justifica las decisiones que tomas en la corrección o curación de los datos.

Ten en cuenta métodos como el (*in place*) `is.na()`, y que puedes vectorizar una negación lógica con `np.logical_not()`. También te puede ser útil el método `pandas.isin()`.

(1.5 puntos) EG

e) Encuentra los valores perdidos en casos y fallecimientos. Imputa los valores perdidos por el valor más frecuente (moda) de cada columna.

(1 punto) NM

[21]:
```
# Respuesta
```

4.1 Ejercicio 2

En este ejercicio trabajaremos en el conjunto de datos derivado del ejercicio anterior.

a) Calcula el valor medio, máximo, mínimo y desviación típica del número de casos.

(1 puntos) NM

[26]:
```
# Respuesta
```

b) Calcula el total de fallecidos por continente anotados en la fecha de 30 de abril del 2020.

(1 punto) NM

[28]:
```
# Respuesta
```

c) Calcula proporción de cases por cada mil habitantes para todo el conjunto de datos. Encuentra la fecha para la que esta proporción és máxima para cada pais.

(1.5 puntos) NM

[30]:
```
# Respuesta
```

d) Crea un nuevo atributo (`estrato`) que contenga la discretización de la variable de casos diarios en 5 segmentos con nombres `Very Low` (de 0 hasta a 10 casos/día), `Low` (de 10 hasta 50 casos/día), `Medium` (de 50 hasta los 100 casos/día), `High` (de 100 hasta los 1000 casos/día), `Very High` (mayores de 1000 casos/día). Lista los paises que estan en el segmento de `Very High` a fecha de 1 de Mayo del 2020.

(1 punto) NM

[32]:
```
# Respuesta
```

4.1.1 Ejercicio Opcional

El fichero de datos de este notebook contiene marcas temporales. Pandas dispone de infraestructura para contener series temporales. El objeto base para anotar marcas de tiempo en Pandas es `Timestamp`. Estudia la documentación de pandas `Series`, `Timestamp` y `to_datetime` para añadir una marca de tiempo al dataframe de este notebok.

Selecciona los 7 paises con mayor población en el mundo, representa gráficamente la evolución temporal de los casos y fallecimientos para cada uno de los 7 paises.

EI

[35]:
```
# Respuesta
```