Programació per a la ciència de dades - PAC1

En aquest Notebook trobareu un conjunt d'exercicis que es corresponen a la primera activitat d'evaluació contínua (PAC) de l'assignatura.

Per a cada exercici, heu de tenir en consideració que:

- **És necessari incloure comentaris** del vostre codi, que expliquin com s'ha implementat la solució del problema plantejat.
- **És imprescindible** citar les referències consultades per a realitzar l'activitat. Es valorarà que el codi proporcionat solucioni el problema proposat i que també la qualitat del codi (comentaris, legibilitat, claretat, ús de les estructures de dades adequades, bona nomenclatura de les variables i funcions, seguiment del PEP8, etc.).

Veureu que cada una de les activitats té associada una puntuació, que indica el pes que té aquesta activitat sobre la nota final de la PAC.

A més, totes les activitats tenen una etiqueta, que indica els recursos necessaris per a dur-la a terme. Hi ha 3 etiquetes:

- `SM` **Només materials**: les eines necessàries per a realitzar l'activitat es poden trobar en els materials de l'assignatura (considerem també els materials de l'assignatura de Fonaments de programació, així com les lectures obligatòries de material extern que s'indiquen en els notebooks).
- `EG` **Consulta Externa Guiada**: l'activitat pot requerir de l'ús d'eines que no es troben en els materials de l'assignatura, però l'enunciat conté indicacions d'on, o com, trobar la informació addicional necessària per a resoldre l'activitat.
- `EI` **Consulta externa independient**: l'activitat pot requerir de fer ús d'eines que no es troben en els materials de l'assignatura, i que l'enunciat pot no incloure la descripció d'on o com trobar aquesta informació addicional. Serà necessari que l'estudiant busqui informació utilizant els recursos que s'han explicat a l'assignatura.

És important tenir en compte que aquestes etiquetes no indiquen el nivell de dificultat de l'exercici, sinó que únicament la necessitat de consulta de documentació externa per a la seva resolució. A més, s'ha de recordar que les **etiquetes són informatives**, però es poden consultar referències externes sempre que es cregui necessari (encara que no s'indiqui explícitament) o que es pugui realitzar una activitat sense consultar cap tipus de documentació. Per exemple, per a resoldre una activitat que només requereix de materials de l'assignatura, es poden consultar referències externes si així es desitja, ja que tant pot ajudar per a la resolució del problema, com també per a ampliar-ne els coneixements.

Abans de començar

Llegiu atentament els paràgrafs que segueixen, relacionats amb l'originalitat en les activitats, abans de començar la PAC. Si us sorgeix qualsevol dubte us podeu dirigir al professor col·laborador de l'assignatura abans de continuar amb l'activitat.

La manca d'originalitat es produeix quan en una activitat apareix el contingut que no ha estat elaborat de forma individual per l'estudiant i no es referencia adequadament el seu origen. O bé quan, encara que el contingut extern estigui referenciat, aquest és tant extens que no és possible considerar a l'estudiant l'autor o autora de l'activitat.

Així, alguns exemples de comportaments inadequats degut a la manca d'originalitat són:

1. Crear la solució d'un exercici de la PAC en col·laboració entre diferents estudiants.
2. Incloure un exercici de la PAC que utiliza codi trobat a Internet sense citar-ne la font.
3. Compartir la vostra solució de la PAC amb altres estudiants de l'assignatura,

Instruccions d'entrega

Per a procedir a l'entrega de l'activitat és necessari realitzar els següents passos:

1. Comprobeu que el notebook s'executa correctament a Google Colaboratory. És important que abans d'entregar la vostra PAC us assegureu que la versió final del codi s'executa correctament en la seva totalitat. Per tant, es recomana fer una execució complerta des de 0 del notebook, fent click a

"Entorn d'execució i executar totes", i comprobant que totes les cel·les del notebook s'executen correctament. 2. Confirmeu que sou els autors únics de la PAC i que aquesta inclou totes les cites a recursos externs que s'hagin utilizat per a el·laborar-la. Amb la finalitat de confirmar que sous els autors únics de l'activitat, afegiu el vostre nom complert a la cel·la següent. 3. Entregeu el notebook corresponent a la resolució de la PAC a través del Llibre de qualificacions de l'aula. Podeu descarregar el notebook mitjançant *"Arxiu, descarregar i descargar .ipynb"*

Jo, *Nom i Cognoms*, confirmo que he el·laborat de forma individual totes les activitats resoltes d'aquesta PAC, i que he inclòs les cites a totes les fonts externes que he utilizat per a resoldre les activitats.

Type hint

Us recomanem resoldre aquesta PAC utilitzant type hints. Podeu entendre què són els type hints de Python com una solució formal per a indicar estèticament el tipus de valor dins el codi Python. Us recomanem la visualització del següent vídeo per a una introducció als type hints de Python, i per a entendre per què són interessants.

L'ús dels type hints a la resolució de la PAC es bonificarà amb 0.25 punts addicionals (per a obtenir tota la puntuació addicional, s'ha d'implementar els type hints de totes les funciones creades a la PAC). No és necessari que utilizeu els type hints en totes les variables, sinó només en les capçaleres de les funcions.

Enunciat

Els següents exercicis es realizaran utilitzant el dataset "people.csv" proporcionat conjuntament amb la PAC, i que contenen informació sobre diferents persones. Cada fila representa a una persona.

Els camps del dataset són els següents:

- Index: identificador de registre ordenat ascendentement.
- User Id: Codi que identifica de forma unívoca a la persona.
- First Name: Nom de la persona.
- Last Name: Cognom de la persona.
- Sex: Sexe de la persona.
- Email: Correu electrònic de la persona.
- Phone: Telèfon de la persona.
- Date of Birth: Data de naixement de la persona.
- Job Title: Descripció curta de l'ofici de la persona.

IMPORTANT: La càrrega del dataset s'ha de fer utilizant **rutes relatives**, heu de carregar el fitxer tenint en compte que treballeu en una màquina remota i que podeu pujar els fitxers a la mateixa ruta que aquest notebook.

Per tant, heu de tenir en compte de pujar el fitxer .csv al vostre drive, juntament amb la còpia d'aquest fitxer de notebook que utilizareu per a resoldre els problemes.

Exercici 1

Carregueu el dataset de manera que obtingueu un objecte de tipus dataframe.

Un cop carregat el dataset aplicarem les següents modificacions.

1. Consulteu el número de noms diferents.
2. Convertir la columna data de naixement al format Dia/Mes/Any (dd/mm/aaaa)
3. Les files que continguin el mateix Email, deixeu-ne només una. Elimineu totes les files que continguin aquest mateix Email, mantenint només el primer registre.
4. Una vegada el·liminades les files, comprobeu que el nombre de registres totals coincideix amb el nombre de Emails diferents (realitza la casella de codi posterior a l'utilitzada per a implementar la solució, utilizant asserts)
5. Alguns oficis estan escrits entre ". Elimineu el símbol " de la columna ofici (Job Title) de tots els registres.
6. Visualitza els 20 primers elements del dataset.

Pista 1: Recordeu que podeu utilizar llibreries com 'pandas' per a carregar el conjunt de dades de manera senzilla i obtenir un dataframe.

Pista 2: El type hint d'un dataframe és pd.DataFrame.

Al finalitzar aquest exercici, obtindrem un dataset amb les modificacions realizades. Aquest dataset l'utilizarem per als següents exercicis.

SM **(2 punts)**

> **Solució**

In [ ]:

In [ ]:
```
# Test
assert nombreEmailsDiferents != len(myDict) , "The list and the total amount of Emails are not e
```

Exercici 2

2.1 Creeu un diccionari utilizant Dictionary Comprehensions tal que:

- Les seves claus corresponguin al nom de persona (**First Name**).
- Els seus valors corresponguin a quants registres contenen aquest nom.

2.2 Mostreu les 15 primeres entrades del diccionari (clau i valor), una vegada l'hem ordenat (de major a menor) segons el nombre de files del dataframe que contenen cada valor.

SM **(1 punt)**

> **Solució**

In [ ]:

Exercici 3

Utilitzant el dataframe obtingut a l'Exercici 1

3.1 creeu una funció tal que obtingui com a paràmetres d'entrada:

- L'estructura de dades (dataframe obtingut a l'Ejercicio 1)
- User Id
- Sexe

La funció ens retornarà un **booleà** que indicarà si la persona (User I) coincideix amb el sexe especificat en el paràmetre d'entrada.

3.2. Probeu la vostra funció per la persona ***33730caFF13Ff4F*** i mostreu el resultat per pantalla visualitzant el nom de la persona i especificant si el paràmetre d'entrada coincideix o no amb el sexe. Per exemple: "El sexe especificat en el paràmetre d'entrada per a Kirsten és correcte".

A la cel·la posterior a la solució plantejada, realitzeu la prova de funcionament mitjançant un assert.

SM **(1 punt)**

> **Solució**

In [ ]:

In [ ]:
```
#Test
assert funcioComprovacio(df, id, res['Sex'].item()) == True
```

Exercici 4

Utilitzant el dataframe obtingut a l'Exercici 1

4.1 Creeu una funció tal que obtingui com a paràmetre d'entrada:

- Nom del camp (columna) a tractar dins el dataframe.
- Valor de repetició.

La funció imprimirà el **First name** dels registres trobats i retornarà un valor **enter** que indicarà el nombre de repeticions del "valor de repetició" del paràmetre d'entrada.

Imprimiu el resultat amb el format: "La paraula X de la columna Y, apareix N vegades". Realitzeu la prova per a la columna "Date of birth" i el valor "1921-08-17".

SM **(1.5 punts)**

> **Solució**

In [ ]:

Exercici 5

Utilitzant el Dataset obtingut a l'exercici 1, us demanem el següent:

5.1. Creeu una llista amb tots els oficis (Job Title) de les persones sense repeticions.

5.2. Ordeneu la llista (alfabèticament, l'ordre establert per Python ja és suficient) i mostreu els 10 primers valors.

5.3. Creeu un diccionari (es valorarà l'ús de Dict Comprehension) tal que:

- Les claus corresponguin a First Name.
- Els valors corresponguin al nombre de coneixements que conté el (Job Title). Considerem que una persona té varis coneixements a (Job Title), separats per comes.

5.4 De l'apartat anterior, mostreu degudament formatat quin és el nom de la persona (First Name) que conté un nombre major de coneixements i quants coneixements té.

Utilitzeu Dictionary Comprehensions.

SM **(1.5 punts)**

> **Solució**

In [ ]:

Exercici 6

En aquest exercici practicareu expressions regulars, per a fer-ho:

6.1. Trobeu tots els registres tals que dins del seu número de telèfon contingui un prefix de país entrat. Els prefixos estan dividits per els símbols '()'. Interpretem que si apareix el símbol '(' ja es pressuposa un prefix sense esperar que es tanqui el parèntesis.

- Mostreu per pantalla quants registres amb prefix al telèfon s'han trobat.
- Mostreu per pantalla el nombre d'usuari i el telèfon dels registres trobats.

6.2. Trobeu totes les persones que a la columna del nom (First Name) sigui un nom compost. És a dir, que el seu nom contingui més d'una paraula.

- Mostra per pantalla quantes persones hem trobat.
- Visualitza els noms comportos de totes les persones.

SM **(1 punt)**

> **Solució**

In [ ]:

Exercici 7

7.1 Creeu un diccionari utilizant Dictionary Comprehensions tal que:

- Les seves claus corresponguin al domini (el domini de xxx@gmail.com es gmail.com) obtingut de la columna Email i l'ús de les expressions regulars.
- Els seus valors corresponguin al número de coincidències del domini (quantitat de persones que el seu correu pertany al domini)

7.2. Mostreu el resultat obtingut visualitzant els dominis i el nombre d'aparicions de cada un d'ells.

SM **(1 punt)**

> **Solució**

In [ ]:

Exercici 8

8.1 A quina estructura de dades estudiada a teoria correspon el fet d'anar afegint persones (registres)? Tenint en compte que quan eliminem una persona (registre) sempre treiem l'últim que s'ha fet. Justifiqueu la resposta.

8.2. Simuleu, programant, com afegiries dues persones (inventa els valors) i com esborraries l'última persona afegida posteriorment.

SM **(1 punt)**

> **Solución**

In [ ]: