Pràctica Pt1.Pandas

Institut: Provençana (L'Hospitalet de Llobregat)
Curs: 2022-2023
Grup: DawBio-2
Mòdul: M14
UF: UF1 – Informàtica mèdica.
Instrument d'avaluació: Pt1.Pandas

Tema: Tractament de dades mèdiques amb Python i Pandas.

Data: 2022-10-14

Notes:

Grups:

- ✔ Aquesta pràctica és fa en grups de dues o tres persones.
- ✔ Registreu el vostre grup i dataset al full de càlcul compartit

Tasques:

- ✔ La pràctica consisteix en formular i contestar 9 preguntes sobre un dataset.
- ✔ El grup pot escollir qualsevol dataset de dades obertes relacionades amb la Biomedicina.
- ✔ S'ha d'utilitzar Pandas, Matplotlib i els continguts d'estadística fets a classe.

Puntuació:

- ✔ Cada pregunta val 1 punt.
- ✔ La nota màxima es ponderarà sobre 10 punts.
- ✔ De la pregunta 4 a la 9 heu d'indicar les preguntes.
- ✔ De totes les preguntes, cal adjuntar el text i el codi que es demana.
- ✔ Per respondre algunes preguntes també caldrà facilitar un gràfic.

Lliurament:

- Heu de pujar el dataset sencer junt amb tots els jupyter notebooks en un arxiu **.zip**.
- Només cal que pugi el .zip un dels membres del grup.
- L'arxiu .zip s'ha de dir «*nom-grup-pt1-pandas.zip*»
- Si el els arxius no segueixen aquest format no es corregiran.
- Pugeu els arxius a la tasca del moodle abans de la data límit.

Datasets

Aquí teniu algunes URLs d'exemple per cercar datasets de portals de dades obertes sobre Biomedicina.
Altres datasets s'altres llocs també són vàlids, però consulteu-lo amb els professors primer.

- https://healthdata.gov/search/type/dataset
- https://www.ncbi.nlm.nih.gov/datasets/
- https://www.who.int/data/collections
- https://data.un.org/
- https://datos.gob.es/en/catalogo?theme_id=salud
- https://www.idescat.cat/dades/

Observacions:

- Escolliu un dataset que us interessi personalment per alguna raó.
- **Si un grup escull un dataset, cap altre grup ja no el pot escollir.**
- Qui primer escriu el seu dataset al full de càlcul compartit, se'l queda.

Format «Tidy»

Format tidy:

1. Cada fila és una observació.
2. Cada columna és una variable.
3. Cada cel·la conté només una dada.

Tutorials de com convertir datasets a format Tidy en Pandas:

- https://www.jeannicholashould.com/tidy-data-in-python.html
- http://shzhangji.com/blog/2017/09/30/pandas-and-tidy-data/
- https://medium.com/@aaronmak/tidying-datasets-in-python-8634f39159bc

És important que aconseguiu que les dades estiguin en format Tidy per poder-les aprofitar per a realitzar estudis.

Preguntes de l'enunciat.

Pregunta 1) Importació del Dataset

1. Explicació del context. Què son aquestes dades? Posar referències.
2. Explicar les columnes que usareu (no cal totes).
   a) Nom
   b) Tipus (string, categorical, data, número enter, decimal ...)
   c) Per a què serveix, si no queda clar amb el nom.
3. Quantes files hi ha?
4. Hi ha NAs? A on?
5. **Resultat final, fitxer Jupyter Notebook amb:**
   a) Text responent les anteriors observacions.

b) Codi font que permeti carregar el CSV en un dataframe i mostri les primeres línies.

Pregunta 2) Arreglar el Dataset.

1. El dataset està en format «tidy»? Justifiqueu la vostra resposta.
2. Si no ho està, poseu-lo en aquest format utilitzant Pandas.
3. **Resultat final, completar el fitxer Jupyter Notebook amb la resposta, i el codi en Pandas que heu usat, si us ha fet falta.**

Pregunta 3) Tractament de valors no disponibles, NaN.

1. Si el fitxer no té valors NaN crea algunes files amb alguns valors NaN.
2. Ara, aplica una d'aquestes dues operacions i justifica el motiu:
   a) Substituïr el valor dels NaN d'una columna per un altre valor. (operació fillna)
   b) Eliminar les files que tinguin algún valor NaN concret. (operació dropna)
3. **Resultat final, completar el fitxer Jupyter Notebook amb la resposta, i el codi en Pandas que heu usat.**

Pregunta 4) Consulta que filtri resultats.

1. Que mostri només algunes de les columnes del dataframe.
2. Que filtri algunes de les files per un o més criteris.
3. **Resultat final, Jupyter Notebook o projecte Python amb el codi.**

Pregunta 5) Consulta que crei un rànquing.

1. És a dir, que ordeni els valors d'una columna i mostri els primers per pantalla.
2. També heu de mostrar un gràfic.
3. **Resultat final, Jupyter Notebook o projecte Python amb el codi i el gràfic.**

Pregunta 6) Consulta que crei almenys una columna calculada.

1. És a dir, que la consulta crei un nou camp depenent d'un altre camp, o calculat a partir d'altres columnes.
2. Exemples:
   a) camp Apte/NoApte depenent de les notes d'alumnes
   b) càlcul imc a partir del pes i l'alçada.
3. **Resultat final, Jupyter Notebook o projecte Python amb el codi.**

Pregunta 7) Consulta amb dades agrupades per un camp de tipus categòric.

1. Si no teniu un camp que es pugui convertir a tipus categòric, haureu de crear-ne un.
2. També heu de mostrar un gràfic de totes les categories.
3. **Resultat final, Jupyter Notebook o projecte Python amb el codi i el gràfic.**

Pregunta 8) Consulta amb dades agrupades per data.

1. És a dir, que si les dades no estan agrupades les haureu d'agrupar per data; ja sigui per any, per mes o per dia.
2. També heu de mostrar un gràfic.
3. **Resultat final, Jupyter Notebook o projecte Python amb el codi i el gràfic.**

Pregunta 9) Separació i fusió de datasets.

1. Tria una de les 2 operacions:
   a) Fes una còpia del dataSet, aconsegueix crear 2 dataSet amb camps i files separats però que comparteixin un camp comú, i després fes el merge.
   b) Si el teu dataSet està desactualitzat o falten dades d'alguns anys i les trobes dades per altres fonts, crea un nou conjunt de dades amb algunes files i/o alguna columna nova. Finalment, fes el merge.
2. **Resultat final, Jupyter Notebook o projecte Python amb el codi.**