Inlämningsuppgift

Inlämningsuppgiften är uppdelad i 3 delar;

1. Numpy
2. Pandas
3. Analys

Varje del har sina egna instruktioner som följer.

> Lämna in en Inlamningsuppgift 2_förnamn_efternamn.ipynb med ersättning av “förnamn” och “efternamn” med ert egna namn.

Betygsättning

Godkänt:	För Godkänt gäller: att klara av hela Numpy delen.
Väl Godkänt:	För Väl Godkänt gäller: allt under Godkänt Svara på VG frågorna under Panda VG delen under Analys

Part 1 - Numpy

I den här delen gäller det att konstruera och manipulera numpy arrays.
Referenshjälp: https://numpy.org/doc/stable/reference/routines.html

Först behöver ni importera numpy.

In [1]:
```
import numpy as np
```

1. Konstruera en array med 10 kolumner och 5 rader av float-tal. Inget tal får vara samma.
2. Konstruera en ny array till med 10 kolumner och 7 rader av float-tal. Den skall inte vara samma som ovan array..
3. Gör om de ovan arrayerna så att de har 5 kolumner men innehåller samma data.
4. Slå ihop de bägge arrayerna till en array.
5. Ändra ordningen på kolumnerna så att de har ordningen kolumn5, kolumn3, kolumn4, kolumn1 och kolumn2.

Part 2 - Panda

I den här delen gäller det att ladda in en panda dataset och undersöka denna.

In [1]:
```
import pandas as pd
```

In [2]:
```
assults = pd.read_csv('assults.csv')
```

Efter att ha läst in datan ta reda på från dataframe nedan saker och skriv ut.

1. Hur många regioner finns med i data?
2. Lista regionerna i bokstavsordning
3. Vad är invånarantalet i varje region
4. Hur många brott begicks i varje region
5. Skapa en dataframe på de områden med de 10 högsta antalet överfall.
6. Lägg till en kolumn i dataframen som visar % antal överfall i området jämfört med invånarantal i området
7. Skapa en ny dataframe som innehåller antalet områden som finns av de olika typerna av urban area. Ledtråd: *groupby*
8. Ta reda på hur många områden hade population men inga överfall. Jämfört totala populationen i dessa med totala poplulationen i deras region.
9. Räkna antalet områden som inte har någon data på överfall eller invånare.
10. **VG**: Skapa en ny dataframe med 10 områderna med mer än 5000 invånare, men har de lägsta överfallen per invånare.
11. **VG**: Skapa en ny dataframe som innehåller % överfall i regionen jämfört med antalet invånare i regionen.
12. **VG**: Lägg till en ny kolumn i dataframen som visar % antal överfall per region

Part 3 - Utforskning

I den här delen gäller det att ladda en valfri panda dataset och undersöka denna.

Dokumentation skrivs i Markdown celler i din Inlämnings.ipynb

Din inlämning kommer värderas på:

- Minst 500 rader, 5 Kolumner i ditt valda dataset
- Fråga och besvara minst 4 frågor om ditt dataset.
- Minst 4 grapher (kanske representerande ovan frågor) med annoteringar så att man kan förstå dem.
- Dokumentation i Markdown celler.
- Ingen plagiat.

Dataset repositories:

UCI repository

Public datasets

Google dataset search

Kaggle datasets

Inspiration:

Analyzing your browser history using Pandas & Seaborn by Kartik

Godawat 2019 State of Javscript Survey Results

2020 Stack Overflow Developer Survey Results

Step-by-step guide

Step 1: Välj dataset

Hitta ett intressant dataset. Det valda datasettet måste vara i CSV format, ha minst 500 rader och 5 kolumner.

Ladda ner datasetet med pandas read_csv funktion och en url (se nedan). Alternativt i inlämningen skicka med datasettet.

```
In [1]:
            import pandas as pd
            url =
'https://raw.githubusercontent.com/cs109/2014_data/master/countries.csv'
            data = pd.read_csv(url)
```

Step 2: Data preparation och Cleaning

Ladda in dataset in i en dataframe med pandas.

Behandla missade data, icke-korrekt data och fel data.

Gör andra steg som behövs (gör om sträng datum till riktiga datum, skapa fler kolumner, slå ihop datasets osv).

Summera hur datasetet ser ut i nuvarande läge i dina markdowns. Storlek, kolumner, kategori typer (qualitative vs. quantitative), kvalitet på data, fång osv.

Step 3: Undersökande analys och visualisering

Undersök data genom analysera (mean, sum, range osv) intressant statistik i numeriska kolumnerna.

Undersök distributions av numeriska kolumner genom histogram osv

Undersök relationen mellan kolumner med scatter plots, bar chars osv.

Notera intressanta upptäckter i din markdown.

Step 4: För **VG** - fråga och svara på 4 intressanta frågor om datan

Fråga minst 4 intressanta frågor om ditt utvalda dataset. Vad kan du göra för analyser på den valda datan?.

Svara på frågorna med antingen resultat från Numpy/Pandas eller skapa plottar med Matplotlib.

Skapa nya kolumner, slå ihop datasets och skapa grupper när det behövs.

Dokumentera din användande av Pandas/Numpy/Matplotlib funktioner och vad de gör i din markdown.

Step 5: Summera dina slutsatser & Skriv ett avslut.

Skriv en summering vad du lärt dig av din analys..

Inkludera intressanta insikter och grafer från föregående sektioner.

För **VG** - Ge dina tankar på vad man kan göra i framtiden inom samma område.

Länkar till Resources du funnit användbara under uppgiften.