DATA TRUTH AND JOINING

There are three files in your Brightspace folder. All of these were supplied after the Irish General Election in 2016. In these elections, the country is divided into constituencies. Each constituency is represented by a specific number of TDs. These elections use proportional representation. A candidate is elected when he / she reaches the constituency quota, or is not eliminated when there is still a seat to be filled. Details of who won is given, however, you will need to have some understanding of the data to be able to generate correct charts. Discuss the files with others in the class.

Significantly, the data is not normalized or cleaned.

The first **candidate** file describes every candidate who went forward for election in every constituency. A candidate is allowed to go for election in more than one constituency. A candidate may or may not be a member of a party. Depending on your final goal, you may not need all attributes.

https://data.gov.ie/dataset/candidate-details-for-general-election-2016

The second **constituency** file describes each constituency in English and as Gaeilge, and has a number of attributes that may or may not be significant, depending on your goals.

https://data.gov.ie/dataset/general-election-2016-constituency-details

The third **count** file records the details of each count. After voting, each candidate's votes are counted. If a candidate has reached the quota, he / she is deemed to be elected. Every candidate's votes for every count are recorded. The count on which the candidate is elected is recorded and that candidate is not involved in any further counts.

https://data.gov.ie/dataset/general-election-2016-count-details

Download the candidate, constituency and count details files from Brightspace. This exercise will merge the three files and plot graphs from it in the following steps.

1) Using the candidate csv, load the data into a data frame called **canddf** (Please note, because of the accents (e.g. Eamon Ó Cuiv), you will need to use a latin1 file encoding). Check if you need to change anything else.
   a) Display the Surname and First.Name of candidates from the Wexford constituency.
   b) How many candidates are there in the Laois constituency?
   c) How many constituencies are there?
2) Download the Constituency details file and load it into a dataframe called **constdf**.
   a) How many constituencies are there in this source?
   b) Assuming the constituency source is correct, look at the data and make a decision about any inconsistencies.
3) Download the count file and load the data into a data frame called **countdf**.
   a) How many candidates are there in the Laois constituency according to countdf?

   b) Check the constituency data for consistency with the new count data. If you find inconsistencies, make a decision on how you will handle them.
   c) Check the candidate data for consistency with the count data.
4) Merge the data sources as you see fit, to end up with a new data frame **df**
   a) Only retain columns: Candidate.First.Name, Candidate.surname, Constituency.Name, Constituency.Number, Count.Number, Gender, Number.Of.Candidates, Number.of.Seats, Party.Abbreviation, Party, Quota, Result, Seats.in.Constituency, Total.Votes, Votes
   b) Generate a new dataframe, **edf,** of elected candidates, with each elected candidate now appearing only once. This dataframe should include the columns Candidate.First.Name, Candidate.surname, Constituency.Name, Gender, Party, Party.Abbreviation
5) **Write a function** that takes a parameter of the constituency name and returns a dataframe with the candidate's first and surnames, the constituency name, gender, party and party abbreviation for all elected candidates in that constituency.
6) Create a plot to show the solutions to the following:
   a) In a constituency (pick one), show the spread of elected candidates across the parties.
   b) Over the whole country, show the spread of elected candidates across the parties.