Final examination

31/08/2017

Download data set for this examination (**Ctrl + S** to Save): http://save.duyet.net/xYJUSl
Description: Open Data The City of Edinburgh Council about Health (GP Practices.csv).

Question:

1. Data cleaning
    a. Remove these columns: **"Telephone Number", "CHP Name", "CHP Code", "Practice Code".**
    b. Cleaning and Standardizing *"Practice Type"* (e.g. **17J**, **2C**).
    c. Formatting the **"Postcode"** as *"XXX-XXX"* (e.g. **EH5-3AH**).
    d. Cleaning and Standardizing **"Public"** column (Y = Yes, N = No). Filling all missing value by **"N".**
    e. Rename the column **"Practice Code"** to **"PCode"**, and **"Practice List Size"** to **"PSize"**.
    f. Replace any negative values in "**PSize**" with the column average.
    g. In the **"Practice Name"** column, standardise the strings so that only the first letter is uppercase (e.g. "*LEVEN MEDICAL PRACTICE*" should become "*Leven Medical Practice*".)
    h. The "**Month_Year"** column would be better as two separate columns! Split each string on the underscore delimiter **_** to give two new columns with the correct values.
2. Data exploration
    a. Select the data in rows **[3, 4, 8, 20-30]** and in columns **["Practice Name", "Telephone Number", "Practice Type", "Public"]**.
    b. Figure out distribution of **"CHP Code"** (count values for each "CHP Code").
    c. List all unique **"Postcode"** values having **"PSize" > 10000**.
    d. Calculate the mean **"Practice List Size"** for each different **"CHP Name"** in dataframe.
    e. Sort dataframe first by the values in the **"PSize"** in descending order, then by the value in the **"PCode"**' column in ascending order.
    f. Select the rows that **"Postcode"** start with *"EH3"*
    g. Save cleaned file to **data_final.xls**

_____ THE END _____