Coursework Part 2 – Pumpkins summary

**Information about the data you will use:**

For this challenge, you will each be given a unique version of a dataset named pumpkins.csv. This dataset contains modified information about the competitive growing of giant pumpkins from the official website of the [Global Pumpkin Commonwealth](http://www.bigpumpkins.com/ViewArticle.asp?id=132). The rows of this dataset are all entries of pumpkin competitions across the world from 2013 to 2021. The columns key is below but remember **not all are relevant to this task**

* id - the year and type of the competition
* place - the place and ranking in the competition
* weight\_lbs - the weight of the pumpkin in pounds
* grower\_name - name of the grower
* city - city pumpkin was grown in
* state\_prov - state or province of grower
* country - country of grower
* gpc\_site - great pumpkin commonwealth site
* seed\_mother - seed mother
* pollinator\_father - father
* ott - over the top inches - for estimating weight
* est\_weight - estimated weight in lbs
* pct\_chart - percentage on chart
* variety - pumpkin variety

**What you should do with the data:**

For this part of your coursework, you will write an R or Python script that will read in the dataset provided to you, modify, summarise and extract information from the data set, generate some plots and then save the plots and modified data set to your computer. You must then create a github repository for this coursework task and upload your script along with accompanying files (described below) to the repository. You will submit the URL of your github repository alongside a zip folder download of your repository.

You may use any R packages or Python modules you wish to to complete this challenge. You will submit a separate text file that lists the modules or packages that your script requires to run. You must load packages / import modules in your script but you can assume they are already installed and do not need to include code to install them in your coursework.

Please bear in mind that part of your mark for this coursework will come from the outcome of running your script on a dataset that is similar, but different to the ones you were each given – the test dataset will have the same column names but information contained within the columns will differ. You will therefore need to ensure that your script can be run from start to finish without generating errors.

Make sure to comment all of your code thoroughly so that it is clear what each line of code is meant to achieve. Your script must achieve the following:

1. Read in the pumpkins.csv data.
2. Identify the heaviest pumpkin grown in any of the competitions – what variety was it? Where was it from and when was it grown?
3. Write a function to change the weight in pounds (lbs) to kilograms (kg) and use this function to create a new column in your data set called weight\_kg.
4. Create another new column in your data set called weight\_class. This column should be a factor with three levels: light, medium and heavy and should be based on the weight of the pumpkins. You can choose the thresholds for these three groups yourself.
5. Plot the relationship between the estimated weight and actual weight of the pumpkins. This can be in either lbs or kg but remember to ensure that **both axes** **are in the same units**. Colour the points in your plot based on the weight\_class column that you created in question 4. Ensure the plot has appropriate axis labels and is clear and well presented. Save this plot to your computer.
6. Filter the data to include only pumpkins from three countries of your choosing and save this filtered data set to your computer in csv format.
7. Summarise your filtered dataset from question 6:
   1. Identify the mean weight of pumpkins for each of your three countries. Which country had the highest mean pumpkin weight?
   2. Identify the mean weight for each variety of pumpkin for each of your countries. Which variety in which country had the lowest mean weight?
8. Using your filtered data set from question 6, plot pumpkin weight distributions (in either lbs or kg) for your three countries as a boxplot. Ensure the plot has appropriate axis labels and is clear and well presented. Save this boxplot to your computer.
9. Redraw your plot from question 8 as a facet plot showing the data from each variety of pumpkin as a separate sub-plot. Save this plot to your computer.

**Your github repository should include:**

1. A single R or Python script named script.R or script.py, containing the code you used to answer the questions above.
2. A text file named requirements.txt containing a list of the R packages or python modules that need to be installed for your code to run.
3. A word document named output.docx containing the answers to questions above and the three plots you generated in questions 5,8 and 9. Include a few sentences summarising what each of the plots shows.
4. A csv file containing the filtered data set that you saved to your computer in question 6 named pumpkins\_filtered.csv
5. Your completed R and Python notebooks from part 1 of your coursework

You can create a zip download of your github repository once you have completed the coursework and uploaded all your files by clicking the green ‘code’ button at the top right-hand side of your repository and then selecting ‘Download ZIP’. Please upload this zip file, along with the URL of your github repository (which can be entered as text) to the Moodle Coursework 1 submission box if you completed this challenge in R or Coursework 2 submission box if you completed it in Python.