General Guidelines

Important:

* Throughout this problem set and future problem sets**, we will be using Python 3.7**. Therefore, if you have Python 3.8 and above, please downgrade it to 3.7! New features in Python 3.8 and above might result in an error during the grading process (for example, the walrus operator ':=' will not be recognize in Python 3.7). Furthermore, many libraries for data analysis are not updated to 3.8 or 3.9 yet, thus, it makes sense to keep using 3.7 for now.

* **Unless specified, you do not have to import any external libraries**.

* **Unspecified libraries will not be taken into account** and therefore will be considered as incorrect answers.

* Whenever a question asks you to write something inside a function, **please do not change the name of the function and the parameters inside it**. You can, of course, make other functions and apply it inside the body of the answer function.

For this problem set, you can use the following libraries:

pandas, numpy, matplotlib, seaborn, statsmodels, json, as well as any Python buit-in libraries (libraries that you do not have to install, they come with Python when you install it)

PROBLEM:

NOTE:

“groups of foods” here means ‘fgroup’ in the dataframe not “food”

You must output specific answers on the console screen, outputting an entire table/figure and eyeballing the answer will not be graded as a correct answer (unless specified)

The database you are working with is a fraction of food nutrient information provided by the US Department of Agriculture (USDA). Most of the data has been altered, therefore, do not use this database for nutritional guidance!

Q1. Which manufacturer has the smallest number of **different** **groups of foods** sent in for analysis? How many groups of foods exactly did the manufacturer send in?

Q2. Each food has a certain number of nutrients, it could be 1, 2 or 3. Assume that nutrients are separated by a comma “,” in your database (for example: Vitamin E, Vitamin C). How many foods are there in your database that have **at least** 2 different nutrients.

Hint: the word “added” or “total” is not a nutrient, for example: “Vitamin E, added” means there is only 1 nutrient, not 2 !!! To make the exercise simpler, these are the only 2 exceptions, other entries are counted as nutrient (even though in real life it is not). For example, “Fatty acids, total saturated” are counted 2 nutrients.

Q3. Which employee (entryById) has the highest amount of entry made in terms of the number of foods analyzed? (Essentially this means which employee is the most productive, he/she should be the one that analyzes the most food samples sent into the organization.). How many different **groups of foods** did this employee had made an entry for?

Q4. Data in the entryById that is not an int (e.g empty or blank or None) should be relabeled as "Anonymous". After relabeling them it seems that the "Anonymous" entry is quite concerning. It is not exactly just a small proportion of the total entry. What is the percentage of the anonymous entries over the total entries?

Q5. You want to investigate if there is a correlation between the number of anonymous entries and certain food groups. Perhaps for some particular **groups of foods**, anonymous entries are more frequent. What **group of food** has the highest frequency of anonymous entries? What **group of food** has the lowest frequency of anonymous entries? Conduct a Chi-square test to check if **groups of food** and the anonymity of entries are independent or correlated, please use all **groups of food**, not a few selective ones. Use the function in the link below to conduct the Chi-square test, you can use Scipy libraries instead of statsmodels library. However, please make sure that you understand the difference between them before using as Scipy by default will result in a different p-value.

For the Chi-square test, you are only required to output the statistic and p-value.

https://www.statsmodels.org/stable/generated/statsmodels.stats.contingency_tables.Table.test_nominal_association.html

Hint: If you have forgotten what Chi-square test is, please check this book https://hds.hebis.de/ubks/Record/HEB451241428, which is available in digital form at the Uni Kassel's library.

Q6. The column entryDate shows the date in which a specific food is submitted to USDA for examination. You are now interested in knowing the distribution of food submission. Perhaps, there are certain months where there are a lot of foods being submitted to USDA for the last 20 years (from 2002-2022). Create a horizontal bar chart that shows the frequency distribution of entryDate by **month** of this database. By month means there should be only 12 bars, each representing a month from January to December.
The bar chart should list the frequency in decending order (e.g: https://www.guru99.com/images/r_programming/032918_1002_WhatisRProg1.png).

Hint: You should be able to finish this question without referring to external resources online by treating the entryDate as string data. The column entryDate is already cleaned! All entries will have the same format which is yyyy-dd-mm. But if you are interested, you can take a look at *datetime* datatype in pandas.

Q7. Make the following function:

def food_for_nutrient(lookup_nutrient, dataframe=default_value):
"""
lookup_nutrient: the name of the nutrient that you want to look up in your database. If
lookup_nutrient does not exist in your database, return "No nutrient found" instead.
dataframe (pandas.DataFrame): the name of the dataframe which we use for looking up, in this
case, it is your database as default_value.

```
return (str): the name of the food that has the highest amount of the lookup_nutrient
"""
```

Use the above function, find out which food has the most vitamin K (Potassium, K)