Exam for Machine Learning Python Lab

Consider the file provided with the assignment and perform the tasks described below.
The solution must be produced as a Python Notebook, assuming that the dataset is in the same folder as the notebook. The notebook must include appropriate comments and must operate as follows:

1. Upload the file `Online-Retail-France.xlsx`. It is a MS Excel file, you can read it with the Pandas function `read_excel`, show the size and a small portion of its content **1pt**

2. It is a transactional database where the role of *transaction identifier* is played by the column `InvoiceNo` and the items are in the column `Description`. Print the number of unique `Description` values ... **1pt**

3. Some descriptions represent the same item but have different leading or trailing spaces, therefore they must be made uniform with the Pandas function `str.strip()` Print the number of unique `Description` values after this cleaning **1pt**

4. Some rows may not have an `InvoiceNo` and must be removed, because they cannot be used. Check if there are such that rows and in case remove them. Inspect the effect of this cleaning. **1pt**

5. Some `InvoiceNo` start with a `C`. They are "credit transactions" and must be removed. Inspect the effect of this cleaning. **1pt**

6. Several transactions include the item `POSTAGE`, which represents the mailing expenses. In this analysis we are not interested in it, therefore the rows with `POSTAGE` will be removed. Inspect the effect of this cleaning. **1pt**

7. After the cleanup, we need to consolidate the items into *one transaction per row* with products one–hot–encoded. To do so, group by `InvoiceNo` and `Description` computing a sum on `Quantity`, use the Pandas `unstack` function to move the items from rows to columns, reset the index, fill the missing with zero, store the result in a new dataframe `basket` and inspect it. **2pt**

8. There are a lot of zeros in the data but we also need to convert to True the positive values and to False the non–positive values. Inspect the result of this transformation and verify the correctness. **1pt**

9. find the maximum value of `min_support` such that the number of rules generated from the frequent itemsets with *lift* not less than 1 is at least 20. Show the value obtained for `min_support` and show the rules.
   Hint: use a loop with an initial value `min_support=1` and decrease it in steps -0.01
   Hint: In `apriori` set the parameter `use_colnames=True`. **2pt**

10. Generate the rules with `association_rules` using `metric=lift` and `min_threshold=1`. **4pt**

11. In order to scatter–plot some information about the rules, it is better to sort them according to some metrics.
    We will sort on descending `lift` and `confidence`, then do a scatter plot of them. **1pt**

..... *see next page*

Quality of the code **4pt**

- Include appropriate comments with reference to the numbered requirements
- Useless cells, pieces of code and non-required output will be penalised
- Remove the code you use for testing and inspecting the variables during the development
- Naming style of variables must be uniform and in English
- Bad indentation and messy code will be penalised
- Non generalised solution, such as three sequential statements with the same kind of operation instead of a loop, will be penalised

Additional directions: the assignments not compliant with the rules below will not be considered:

- The *notebook name* must be in lowercase according to the pattern `workplacecode_youremailusername.ipynb`
  E.G. if your email is `mario.rossi45@studio.unibo.it` and you are sitting in the workplace `lab4_023`, the notebook filename will be `lab4_023_mario.rossi45.ipynb`
- The solution must directly access the data in the default folder.
- Check *carefully* that your file is correctly stored as a python notebook. Upload the notebook only to `http://virtuale.unibo.it` in the activity specified by the teacher, any other way of submitting the notebook will be ignored.
- Cooperative work will be heavily sanctioned both for the giver and the receiver of the copy.
- The candidate must use only the computers of the lab and can access only the official documentation of Python and the other libraries used. Any other device is not allowed for any reason.