CSV

The csv (Purchase.csv) was created in the following steps from the script (csv\_creation.py):

1. The python script imports the libraries random, math, names, csv and datetime, all of which are used in the creation of our csv.
2. Every date between 1/01/2000 and 31/12/2016 is generated as a datetime object in the format yyyy-mm-dd and stored in a list. Code was used from <https://stackoverflow.com/questions/7274267/print-all-day-dates-between-two-dates>.
3. A dictionary of prices for each product is created which ensures that the prices match the prices from our html file. Even though prices aren’t needed for the csv, this is still done because we’ve chosen our shipping prices to be 10% of the product price thus, a dictionary of shipping prices is created from this product prices dictionary.
4. Next a list of 200 random names is created from the names package. These names are then concatenated by taking their surname followed by the initial of their first name in all lower case to create 200 usernames (e.g. James Bailey becomes baileyj).
5. Then a header is created to be added to our csv later.
6. Next the 2d data is created generating the following list 160000 times and appending each list to another list.
   1. A random product id is chosen with random.choice.
   2. A random customer is chosen at random from our list of 200 customers.
   3. A random quantity of product purchased using the beta distribution from the match package with alpha and beta equal to 2. This result is added to 1 to avoid quantities of zero and floored to avoid non integer quantities.
   4. A random date is chosen from our list of dates.
7. The corresponding shipping price for the relevant product and quantity is taken.
8. Note that 160000 transactions was chosen as the number of transactions as is approximately the number of transactions that would occur if 200 customers made 50 transactions a year for 16 years.
9. The csv is then sorted by date which is made easier by our date objects still being of type datetime. The dates are then changed to the format dd/mm/yyyy from yyyy-mm-dd.
10. Every transaction is gone through to ensure that certain products only had a quantity of 1 (e.g. TVs, video game consoles). This is our “interesting” signal.
11. The header is then inserted to be the first row of our 2d data.
12. The 2d data is written to a csv called Purchase.csv using the csv package. Code was used from ## Help from <https://stackoverflow.com/questions/44691524/write-a-2d-array-to-a-csv-file-with-delimiter>.