Yi Zhou

HW1

Jan. 27th. 2022

1. Commend: $ zcat arcos_all_washpost.tsv.gz | head -1 > output1
   Output: REPORTER_DEA_NO REPORTER_BUS_ACT REPORTER_NAME
   REPORTER_ADDL_CO_INFO REPORTER_ADDRESS1 REPORTER_ADDRESS2
   REPORTER_CITYREPORTER_STATE REPORTER_ZIP REPORTER_COUNTY
   BUYER_DEA_NO BUYER_BUS_ACT BUYER_NAME
   BUYER_ADDL_CO_INFO BUYER_ADDRESS1 BUYER_ADDRESS2
   BUYER_CITY BUYER_STATE BUYER_ZIP BUYER_COUNTY
   TRANSACTION_CODE DRUG_CODE NDC_NO DRUG_NAME
   QUANTITY UNIT ACTION_INDICATOR ORDER_FORM_NO
   CORRECTION_NO STRENGTH TRANSACTION_DATE
   CALC_BASE_WT_IN_GM DOSAGE_UNIT TRANSACTION_ID
   Product_Name Ingredient_Name Measure
   MME_Conversion_Factor Combined_Labeler_Name
   Revised_Company_Name Reporter_family dos_str

2. Commend: $ zcat arcos_all_washpost.tsv.gz | wc -l >wc.txt
   Output: 178598027

3. To get a random 5000 line I used:
   Commend: $ zcat arcos_all_washpost.tsv.gz | shuf -n5000 > output2
   Output is too long not showing here
   To get the file that only have the transaction dates:
   Commend: $ cat output2.txt | awk -F '\t' '{print $31}' > output3.txt
   To calculate the estimated proportion, I wrote python code as follows:

```
result = []

with open('output3.txt', 'r') as f:
    for line in f:
        newLine = int(line)
        year = newLine % 10000
        result.append(year)
    newResult = dict((x,result.count(x)) for x in set(result))

newResult = {x: x * 178598027 / 5000 for x in newResult}

print(newResult)
```

Here is the result:
{2006: 71653528.4324, 2007: 71689248.0378, 2008: 71724967.6432, 2009: 71760687.2486, 2010: 71796406.854, 2011: 71832126.4594, 2012: 71867846.0648}