Mattia Danese
CS 119 - Big Data
Professor Singh
Quiz 2: The Opioid Files

Problem 1

The column names of the opioid dataset are:

- REPORTER_DEA_NO
- REPORTER_BUS_ACT
- REPORTER_NAME
- REPORTER_ADDL_CO_INFO
- REPORTER_ADDRESS1
- REPORTER_ADDRESS2
- REPORTER_CITY
- REPORTER_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

The commands I ran:

- `zcat /comp/119/arcos_all_washpost.tsv.gz | head -1`

Problem 2

The opioid dataset has a total of 178,598,026 rows.

The commands I ran:

- `zcat /comp/119/arcos_all_washpost.tsv.gz | wc`
- The above command returns 178,598,027 total lines, where each line represents one row, though 1 must be subtracted from this output as one of the lines is the header row and this is not an actual row in the dataset.

Problem 3

Total number of rows for 2006: 21860398
Total number of rows for 2007: 23574939
Total number of rows for 2008: 24146453
Total number of rows for 2009: 25360920
Total number of rows for 2010: 26575386
Total number of rows for 2011: 28647123
Total number of rows for 2012: 28432806

The commands I ran:

- `zcat /comp/119/arcos_all_washpost.tsv.gz | shuf -n 5000 > temp.txt`
- I then made a python script that reads through all the lines of `temp.txt`, records how many rows correspond to each year, and then multiplies these proportions by the total number of rows in the opioid dataset

```
f = open("temp.txt")
lines = f.read().split('\n')

count = {
   '2006' : 0, '2007' : 0, '2008' : 0, '2009' : 0, '2010' : 0,
   '2011' : 0, '2012' : 0,
}

for line in lines:
   fields = line.split('\t')

   date = fields[30]
   year = date[-4:]

   count[year] += 1
print(count)

TOTAL_ROWS = 178598026
for year in count.keys():
   print("The estimated number of rows for {} is: {}".format(
      year,
      round((count[year] / 5000) * TOTAL_ROWS)))
```