**PROJECT ON ANALYIS OF EXPENDITURE OF CONGRESS REPRESENTATIVES**

**-------------------------------------------------------------------------------------------**

Members of the Congress and Congressional offices receive an annual budget to spend on staff, supplies, transportation and other expenses. Each quarter, representatives report the recipients of their expenditure. ProPublica compiles these reports into research-ready CSV files.

Data for this project was downloaded from the following website:

<https://www.propublica.org/datastore/dataset/house-office-expenditures>

All analysis for this project is based on detailed data (not summary data).

The dataset is described in “expenditures\_readme.txt” file.

NOTE: There is an updated version of the 2015Q2 file and so I have renamed "2015Q2-house-disburse-detail.csv" to

"2015Q2-house-disburse-detail-old.csv".

Then I renamed "2015Q2-house-disburse-detail-updated.csv" to

"2015Q2-house-disburse-detail.csv".

**Part 1: Analysis of total of all payments in the dataset.**

See part1\_analysis.ipynb or part1\_analysis.pdf file.

Total expenditure for all 35 quarters was $5570133308.79.

**Part 2: Calculation of mean and standard deviation of ‘COVERAGE PERIOD’**

“COVERAGE PERIOD” is the difference in days between ‘START DATE’ and “END DATE’ for each payment.

Technical Notes:

1) ’AMOUNT’ column is of type string for the first 31 quarters. Need to convert this to int.

https://stackoverflow.com/questions/42719749/pandas-convert-string-to-int

use pd.to\_numeric(df['ID'], errors='coerce') to convert those values to NaN, note that this will produce a dtype of float

Note: The last 4 quarters had ‘AMOUNT’ column as a float.

2) Adding all the rows of a column in pandas data frame

http://blog.mathandpencil.com/column-and-row-sums

df**.**sum(axis**=**0)

3) Getting low memory.So splitting the file\_list into pieces

4) Getting dtype warning for the last file— 2018Q1

Ref: https://stackoverflow.com/questions/12468179/unicodedecodeerror-utf8-codec-cant-decode-byte-0x9c

Changing the engine from C to Python did the trick for me.

Engine is C:

pd.read\_csv(gdp\_path, sep='\t', engine='c')

'utf-8' codec can't decode byte 0x92 in position 18: invalid start byte

Engine is Python:

pd.read\_csv(gdp\_path, sep='\t', engine='python')

===============================================