CS 649 Big Data
Spring Semester, 2021
Assignment 2

San Diego State University -- This page last updated 3/29/21

Due April 18 11:59 pm

2020 Election Data

You are going to look donation data from the 2020 presidential campaign.

This assignment uses two datasets. A partial dataset for developing and the full dataset. The partial dataset can be downloaded from the assignment page. A zipped version of the full dataset is available on the assignment page. It expands to a 3GB file. The full dataset can be accessed on AWS at s3://rw-cs696-data/P00000001-ALL.csv. The format of the files is described at the end of the document. The partial dataset has two differences from the original. The differences are described at the end of this document.

The data contains information about each donation made to presidential candidates in 2020. The source of the data is:

https://www.fec.gov/data/candidates/president/presidential-map/

Questions

1. How many donations did each candidate have?

2. What was the total amount donated to each candidate?

3. How many unique contributors did each candidate have?

4. What mean and standard deviation of the donations for each candidate.

5. What percentage of the each campaign's donations was done by small contributors, that is donations under $50?

6. Produce a histogram of the donations for the Trump and Biden campaign? The x-axis the amount of the donation and the y-axis the number of donors that gave that amount.

You are to use AWS Spark to answer the first 5 questions. For the 6th question you need to process the data on AWS and download the result so you can use Python plotting tools to produce the histograms.

Instructions

Turn in a jupyter notebook with all the code used with the answer to the questions. The code for problems 1 - 5 should be in a function that you run on AWS. To show that you ran the code

on AWS include in your notebook the AWS CLI export command for each job that you need to run on AWS.

Grading

10 points per problem.

What to turn in

You need to turn in the jupiter notebook and the files that you download from AWS to answer problem 6. Put them in the same directory so the notebook can read the files as local files. Create a zip file of the directory and turn that zipped directory in.

Late Penalty

An assignment turned in 1-7 days late, will lose 5% of the total value of the assignment per day late. The eight day late the penalty will be 40% of the assignment, the ninth day late the penalty will be 60%, after the ninth day late the penalty will be 90%. Once a solution to an assignment has been posted or discussed in class, the assignment will no longer be accepted. Late penalties are always rounded up to the next integer value.

Data Format

Note that some(all?) of the rows in the data set contain an extra column.

Columns

cmte_id - ID of the committee that received the donation. Example: C00285254

cand_id - Candidate ID. Example: "P00013649"

cand_nm - Candidate name. Name is quoted as it contains a coma. Example: "Sanford, Marshall"

contbr_nm - Contributor name. Name is quoted as it contains a coma. Example: "KEITHLEY, BRAD"

contbr_city - Contributor city. Example: "ORANGE BEACH"

contbr_st - Contributor state. Example: "AK"

contbr_zip - Contributor zip. Example: "99501"

contbr_employer - Contributor employer. Example: "BONAPARTE FILMS LLC"

contbr_occupation - Contributor occupation. Example: "CONSULTANT"

contb_receipt_amt - Contributed amount. Example: 1000

contb_receipt_dt - Date contributed. Example: 09-SEP-19

receipt_desc - Often blank. Example: ""

memo_cd - Often blank. Example: ""

memo_text - Often blank. Example: ""

form_tp - Example: "SA17A"

file_num - a unique number assigned to a report and all its associated transactions. Example: "1376946"

tran_id - Example: "AFBC1B0EF531D4CDCBE8"

election_tp - This code indicates the election for which the contribution was made. EYYYY (election plus election year). Options are: (P)rimary, (G)eneral, (O)ther, (C)onvention, (R)unoff, (S)pecial, or (R)ecount. Example: "P2020"

A slightly longer description of the columns can be found at: http://www2.stat.duke.edu/~cr173/Sta102_Sp16/Lab/lab9.html

Partial Dataset Difference

In the original data all text data entires are quoted. In the partial dataset only the text data entires that contain a comma character (,) are quoted.