Exploratory_Work-Group5

October 4, 2020

0.1 Exploratory Data Analysis

5.11 Assignment

Brad Howlett (bth2g)
Eric Larson (rel4yx)
Hanim Song (hs4cf) —

[1]:

```
import os
from pyspark.sql import SparkSession
import pyspark.sql.types as typ
import pyspark.sql.functions as F
from pyspark.sql import DataFrame
from functools import reduce
from pyspark.sql.functions import col, asc
from pyspark.sql import SQLContext
from pyspark.sql.types import StructType
from pyspark.sql.types import StructField
from pyspark.sql.types import StringType, DoubleType, DateType
from pyspark.sql import SparkSession
import pandas as pd
import numpy as np
```

[2]:

```
spark = SparkSession \
    .builder \
    .getOrCreate()

sc = spark.sparkContext
```

0.1.1 Based on data from : https://www.fec.gov/data/browse-data/?tab=bulk-data

This is House/Senate campaign finance data:

[3]:

```
#storing all files in dictionaries for easy reference
f_campaigns = {'2019-2020':'campaign_2020.txt',
               '2017-2018':'campaign_2018.txt',
               '2015-2016':'campaign_2016.txt'}
```

```
#no header provided by FEC to upload
header_row = ['CAND_ID',
              'CAND_NAME',
              'CAND_ICI',
              'PTY_CD',
              'CAND_PTY_AFFILIATION',
              'TTL_RECEIPTS',
              'TRANS_FROM_AUTH',
              'TTL_DISB',
              'TRANS_TO_AUTH',
              'COH_BOP',
              'COH_COP',
              'CAND_CONTRIB',
              'CAND_LOANS',
              'OTHER_LOANS',
              'CAND_LOAN_REPAY',
              'OTHER_LOAN_REPAY',
              'DEBTS_OWED_BY',
              'TTL_INDIV_CONTRIB',
              'CAND_OFFICE_ST',
              'CAND_OFFICE_DISTRICT',
              'SPEC_ELECTION',
              'PRIM_ELECTION',
              'RUN_ELECTION',
              'GEN_ELECTION',
              'GEN_ELECTION_PRECENT',
              'OTHER_POL_CMTE_CONTRIB',
              'POL_PTY_CONTRIB',
              'CVG_END_DT',
              'INDIV_REFUNDS',
              'CMTE_REFUNDS']
```

Create dataframes and combine the three files together for analysis:

[63]:

```
df_temp20 = sc.textFile(f_campaigns['2019-2020']).map(lambda row: [elem for␣
 ↪elem in row.split('|')])
df_temp18 = sc.textFile(f_campaigns['2017-2018']).map(lambda row: [elem for␣
 ↪elem in row.split('|')])
df_temp16 = sc.textFile(f_campaigns['2015-2016']).map(lambda row: [elem for␣
 ↪elem in row.split('|')])

#default to stringtype for ease of loading, then adjust below:
fields = [*[typ.StructField(h[:], typ.StringType(), True) for h in header_row]]
schema = typ.StructType(fields)

df_20 = spark.createDataFrame(df_temp20, schema)
df_18 = spark.createDataFrame(df_temp18, schema)
```

```
df_16 = spark.createDataFrame(df_temp16, schema)

dfs = [df_20, df_18, df_16]

df = reduce(DataFrame.unionAll, dfs)

#casting necessary numeric values:
df = df.withColumn('TTL_RECEIPTS', df['TTL_RECEIPTS'].cast(DoubleType()))
df = df.withColumn('TTL_INDIV_CONTRIB', df['TTL_INDIV_CONTRIB'].
 ↪cast(DoubleType()))
df = df.withColumn('CAND_CONTRIB', df['CAND_CONTRIB'].cast(DoubleType()))
df = df.withColumn('OTHER_POL_CMTE_CONTRIB', df['OTHER_POL_CMTE_CONTRIB'].
 ↪cast(DoubleType()))
df = df.withColumn('POL_PTY_CONTRIB', df['POL_PTY_CONTRIB'].cast(DoubleType()))
```

Number of records:

[51]:

```
df.count()
```

[51]: 7149

Number of columns:

[16]:

```
len(df.columns)
```

[16]: 30

Statistical summary of response variable:

Our statistical summary will be based on whether a candidate won or lost the relevant political race.

We are still gathering and joining that data to this set.

Statistical summary of potential predictor variables:

Total receipts -

[54]:

```
df.select('TTL_RECEIPTS').describe().show()
```

```
+-------+------------------+
|summary|      TTL_RECEIPTS|
+-------+------------------+
|  count|              7149|
|   mean|2266553.4508938333|
| stddev|6.004019782547508E7|
|    min|                0.0|
|    max|      4.824617973E9|
+-------+-------------------+
```

Contributions by individuals -

[55]:

```
df.select('TTL_INDIV_CONTRIB').describe().show()
```

```
+-------+-------------------+
|summary|  TTL_INDIV_CONTRIB|
+-------+-------------------+
|  count|               7149|
|   mean|  3570474.139283814|
| stddev|2.231747534710199E8|
|    min|                0.0|
|    max|    1.8853982587E10|
+-------+-------------------+
```

Contributions by candidates -

[57]:

```
df.select('CAND_CONTRIB').describe().show()
```

```
+-------+-------------------+
|summary|       CAND_CONTRIB|
+-------+-------------------+
|  count|               7149|
|   mean|  621843.8302853543|
| stddev|3.607460293607057E7|
|    min|                0.0|
|    max|      2.831281203E9|
+-------+-------------------+
```

Contributions from party committees -

[64]:

```
df.select('POL_PTY_CONTRIB').describe().show()
```

```
+-------+------------------+
|summary|   POL_PTY_CONTRIB|
+-------+------------------+
|  count|              7149|
|   mean|1594.4136438662752|
| stddev| 37564.26071346217|
|    min|               0.0|
|    max|         3100000.0|
+-------+------------------+
```

Contributions from other political committees -

```
[65]: df.select('OTHER_POL_CMTE_CONTRIB').describe().show()
```

```
+-------+----------------------+
|summary|OTHER_POL_CMTE_CONTRIB|
+-------+----------------------+
|  count|                  7149|
|   mean|     449703.3054748916|
| stddev|  2.2751089972151406E7|
|    min|                   0.0|
|    max|           1.9235003E9|
+-------+----------------------+
```

Candidate status (C = Challenger, O = Open, I = Incumbent) -

```
[61]: #some data cleaning to do for the blanks
      df.groupby('CAND_ICI').count().orderBy('count', ascending = False).show()
```

```
+--------+-----+
|CAND_ICI|count|
+--------+-----+
|       C| 3857|
|       O| 1779|
|       I| 1441|
|        |   72|
+--------+-----+
```

Candidate party affiliation (count) -

```
[41]: df.groupby('CAND_PTY_AFFILIATION').count().orderBy('count', ascending = False).
       ↪show()
```

```
+--------------------+-----+
|CAND_PTY_AFFILIATION|count|
+--------------------+-----+
|                 DEM| 3227|
|                 REP| 3193|
|                 IND|  272|
|                 LIB|  136|
|                 GRE|   55|
|                 NPA|   37|
|                 DFL|   36|
|                 OTH|   35|
|                 NNE|   32|
|                 UNK|   26|
|                  UN|   23|
|                 CON|   14|
|                   W|    9|
|                 NON|    7|
|                 IDP|    5|
|                 NOP|    5|
|                 AMP|    3|
|                 PPY|    3|
|                 SEP|    3|
|                 UNI|    3|
+--------------------+-----+
only showing top 20 rows
```

Candidate state (count) -

[59]:

```
df.groupby('CAND_OFFICE_ST').count().orderBy('count', ascending = False).show()
```

```
+--------------+-----+
|CAND_OFFICE_ST|count|
+--------------+-----+
|            CA|  695|
|            TX|  604|
|            00|  513|
|            FL|  486|
|            NY|  364|
|            PA|  263|
|            GA|  235|
|            IL|  228|
|            NC|  224|
|            MI|  189|
|            OH|  183|
|            VA|  171|
|            AZ|  170|
|            NJ|  170|
|            IN|  148|
|            TN|  147|
|            MD|  147|
|            WA|  132|
|            MN|  129|
|            CO|  125|
+--------------+-----+
only showing top 20 rows
```

[66]:

```
df.select('CAND_NAME',
          'CAND_OFFICE_ST',
          'CAND_PTY_AFFILIATION',
          'CAND_ICI',
          'TTL_RECEIPTS',
          'CAND_CONTRIB',
          'TTL_INDIV_CONTRIB',
          'POL_PTY_CONTRIB',
          'OTHER_POL_CMTE_CONTRIB').show(5)
```

```
+--------------------+--------------+--------------------+--------+------------+
------------+-----------------+---------------+----------------------+
|           CAND_NAME|CAND_OFFICE_ST|CAND_PTY_AFFILIATION|CAND_ICI|TTL_RECEIPTS|
CAND_CONTRIB|TTL_INDIV_CONTRIB|POL_PTY_CONTRIB|OTHER_POL_CMTE_CONTRIB|
+--------------------+--------------+--------------------+--------+------------+
------------+-----------------+---------------+----------------------+
|     YOUNG, DONALD E|            AK|                 REP|       I|  1362383.63|
0.0|        637025.31|            0.0|             584444.63|
|       GALVIN, ALYSE|            AK|                 IND|       C|  2266364.63|
3394.63|        2116292.8|            0.0|              109350.0|
|     AVERHART, JAMES|            AL|                 DEM|       O|    50126.74|
0.0|         23281.74|            0.0|                   0.0|
|    GARDNER, KIANI A|            AL|                 DEM|       O|   118661.85|
764.97|         92896.88|            0.0|               19000.0|
|COLLINS, FREDERIC…|            AL|                 DEM|       O|    62935.42|
56500.0|          5917.12|            0.0|                   0.0|
+--------------------+--------------+--------------------+--------+------------+
------------+-----------------+---------------+----------------------+
only showing top 5 rows
```

[ ]: