IS843 term project:

Introduction:

Voter participation is essential for success of any democratic system. Strong democracies can only exist if the level of voter participation is high. With the increase in racial diversity in the US, the role of race in the US democratic system is drawing more attention of our policy makers in the recent times. While we are aware that there is disparity in the levels of voter participation among different races, the different causes are not so clearly identified. Analysis of relevant data is an important step towards finding answers to such questions.

The voter files:

In the past, the availability of data was a bottleneck to finding answers to such questions. In recent times the amount of data has grown exponentially. Application of big-data techniques has become very important in dealing with such volumes of data. The center for antiracist research (CAR) at BU recently acquired a 3-year subscription to a national voter file for the US. The voter files contain records for every registered voter in the US. A voter files exists for each state. Each file contains geographic, demographic and household information. It also contains the history of voting for each registered voter.

How to access the voter files:

The files have been downloaded to a google cloud bucket to make them easily accessible. The bucket has been made public to everyone in the class. To list the content of the bucket, open a google cloud shell and type:

gsutil -u <project-id> ls gs://car-l2-voter-files/

where project-id is the id of your project. Note don't include <>. You will notice 2 folders named VM2Uniform and VMFiles. You will have to copy the files from the VM2Uniform folder over to a bucket storage under your project.

First create a bucket with an appropriate name under your project. To copy files to your storage bucket, use the following command in the google cloud shell:

gsutil -u <project-id> -m cp gs://car-l2-voter-files/VM2Uniform/VM2Uniform--WY--2021-01-13.zip gs://<NAME-OF-YOUR-OWN-BUCKET>/

where NAME-OF-YOUR-OWN-BUCKET is the name of bucket storage associated with your project. This command will copy the voter file for the state of Wyoming into your bucket storage space (note that unlike regular unix command, there is no . at the end of the command line).

The same procedure can be followed for files for the other states. Note that the copied file is a .zip file. You will have to download the .zip file to your machine and extract the files to your local machine. Each .zip file contains 2 files, the main data set for the state and a dictionary file containing information about the columns of data in the dataset. For example, the extracted file for the State of Wyoming will looks like:

VM2Uniform--WY--2021-01-13.tab
VM2Uniform--WY--2021-01-13_DataDictionary

A word of caution, the files for some states are quite large. For example, the .tab file for the state of California is greater than 50Gb. You need to be aware of the available storage space on your local machine. Teams may want to distribute this task and assign to each team member the responsibility for downloading and extracting files for a few states.

Next, upload the files to your project bucket storage space. Go to Cloud Storage->Browser. Click on your bucket. Click on UPLOAD FILES option to upload the extracted .tab file to your GCP bucket storage.

You are now ready to work with this file in pySpark.

How to read the .tab file in pySpark:

The data in the .tab file is separated by tabs. You can read a .tab file as a .csv file with a delimiter option. The example below shows how to read the file for the state of Wyoming:

```
df = spark.read.format("csv")\
  .option("header", "true")\
  .option("nullValue", "NA")\
  .option("delimiter","\t")\
  .option("inferSchema", "true")\
  .load("gs://is843-team2-project/VM2Uniform--WY--2021-01-13.tab")
```

Note the use of “\t” as a delimiter when reading the file.

What are the contents of a Voter file?

There are 726 columns of data within each file. The accompanying dictionary file is useful for understanding the information contained in the file. The following is the partial output of the schema.

```
df.printSchema()
```

```
root
 |-- SEQUENCE: integer (nullable = true)
 |-- LALVOTERID: string (nullable = true)
 |-- Voters_Active: string (nullable = true)
 |-- Voters_StateVoterID: integer (nullable = true)
 |-- Voters_CountyVoterID: string (nullable = true)
 |-- VoterTelephones_LandlineAreaCode: integer (nullable = true)
 |-- VoterTelephones_Landline7Digit: string (nullable = true)
 |-- VoterTelephones_LandlineFormatted: string (nullable = true)
 |-- VoterTelephones_LandlineUnformatted: long (nullable = true)
 |-- VoterTelephones_LandlineConfidenceCode: integer (nullable = true)
 |-- VoterTelephones_CellPhoneOnly: boolean (nullable = true)
 |-- VoterTelephones_CellPhoneFormatted: string (nullable = true)
 |-- VoterTelephones_CellPhoneUnformatted: long (nullable = true)
 |-- VoterTelephones_CellConfidenceCode: integer (nullable = true)
 |-- Voters_FirstName: string (nullable = true)
 |-- Voters_MiddleName: string (nullable = true)
 |-- Voters_LastName: string (nullable = true)
 |-- Voters_NameSuffix: string (nullable = true)
 |-- Residence_Addresses_AddressLine: string (nullable = true)
 |-- Residence_Addresses_ExtraAddressLine: string (nullable = true)
 |-- Residence_Addresses_City: string (nullable = true)
 |-- Residence_Addresses_State: string (nullable = true)
 |-- Residence_Addresses_Zip: integer (nullable = true)
 |-- Residence_Addresses_ZipPlus4: integer (nullable = true)
 |-- Residence_Addresses_HouseNumber: string (nullable = true)
 |-- Residence_Addresses_PrefixDirection: string (nullable = true)
 |-- Residence_Addresses_StreetName: string (nullable = true)
 |-- Residence_Addresses_Designator: string (nullable = true)
 |-- Residence_Addresses_SuffixDirection: string (nullable = true)
 |-- Residence_Addresses_ApartmentNum: string (nullable = true)
 |-- Residence_Addresses_ApartmentType: string (nullable = true)
```

Many columns have missing data. You will have to decide how to handle the missing data. Since there are many columns of data, teams may want to split up the task and assign a set of columns to each team member to explore. The files contain information at the county level as well as the districts within the county. Since some counties in different states have the same name, doing analysis at the county level across different states poses a problem. The Federal information processing system (FIPS) code is a 5-digit unique code assigned to a county. The FIPS county code by states can be found in the supplied uscounties.csv file in the same public bucket. This file should be used to either insert or replace county names with FIPS codes.

Some questions to consider:

Recent research from the Pew research center shows a widening gap in values between voters supporting different candidates. This includes race, gender, family, immigration and religion. Survey shows that since 2016, there has been a 9% increase in voters who believe that

it is lot more difficult to be black. By contrast, opinions among all voters have changed little on whether women continue to face obstacles that make it harder for them to advance in life and career. Divide also exists on opinions related to immigration and religion. According to the Pew research center, the top 4 voter issues for the 2020 election were economy, health care, supreme court appointment and the coronavirus outbreak. Are the gaps in opinions between the races somehow related to the level of participation in the electoral system? Since voter opinions eventually shape government policies, such survey results are important. Government policies should improve voter participation and lower disparity in the participation levels between the different races. But first, we must understand what factors influence disparity in voter participation. Is access to good schools and education system driving the disparity? Could access to good public transportation improve voter turnout for racial minorities? Is there a connection between access to healthcare and voter turnout? These are just a few of the important questions that researchers have been investigating. More importantly, given the limited resources available, how should the resources be spent to maximize voter turnout among racial minorities.

Answers to such interesting questions can be investigated with our dataset. As an example, Prof. Maxwell Palmer from the political Science department at BU used the dataset to study the effects of car ownership on voter turnout. Use the voter data set to explore voting related question. Since the dataset comes from the center for antiracist research (CAR), use the race related information while exploring the datasets. Include information from other fields such as health care etc. to explore specific questions. To get you started here are a few questions you might wish to explore.

Based on the data in the voter files, can we find answers or explain some of the following questions?

Is there disparity in voter turnout among households based on income level?
How does house ownership and/or housing cost impact voting turnout?
Do counties with higher home ownership have higher voter turnout?

Does location of voting stations have any impact on voter turnout?
Is there a disparity in the number of voting stations between affluent areas and less affluent areas/counties?
Is there a connection between the quality of school district and voter turnout?

What are leading causes for disparity in turnout for voters of different ethnicity?
What are some of the disparities in interests/lifestyle (leisure activity, sports, arts, travel etc.) that impact voter turnout?

Is there disparity in turnout between primary and general election for different ethnicities?

If so, can we find some of the causes for the disparity?

Can we tell from the data what ethnicities are likely to vote by mail?

These questions should be considered as the starting question for investigating specific topics. Teams should not restrict their investigations to only these questions. It is also important not to infer causal relationships while exploring correlations between different entities. For example, finding a positive correlation between household income and voter turnout, is not sufficient to conclude that low household income is the cause for low voter turnout.

References:

1. Driving Turnout: The Effect of Car Ownership of Electoral Participation, Kessner, J.B, Palmer, M, Political Science Research and Methods, 2021.
2. Important issues in the 2020 election: https://www.pewresearch.org/politics/2020/08/13/important-issues-in-the-2020-election/
3. Voters' attitude about race and gender are even more divided than in 2016. https://www.pewresearch.org/politics/2020/09/10/voters-attitudes-about-race-and-gender-are-even-more-divided-than-in-2016/
4. Eviction and Voter turnout: The political consequences of housing instability, Slee, G. and Desmond M., Politics and Society, SAGE, 2021.
5. Equity at the ballot box: Health as a resource for political participation among low-income workers in two United States cities, McGuire, C.M., Gollust S.E., Marco, M.D., Durfee T., Wolfson J. and Caspi C., Frontiers in political science, 2021.
6. When does inequality demobilize? New evidence from the American states, Macdonald, D., Electoral Studies, 2021.