# **Murmuration Data Challenge - Data Scientist**

We have contrived this example based on analyses that Murmuration sometimes conducts for our clients. Please provide the details of your analysis (including your code), as well as any discussion points describing assumptions, limitations, caveats, and next steps using a medium of your choice (e.g., slides, document, Jupyter note­book). Feel free to include any other interesting analytic results that you uncover. Please take as long as you need to finish this exercise, but we are anticipating ~4 hours of work. Please send us the results of the analysis within 72 hours after you receive it. Although these data have been scrubbed to remove the obviously identifying characteristics, they are by no means fully anonymized, *so please delete these files after the completion of the exercise and do not share the data or analysis. Please do not upload the data/analyses or report to github.*

### **Section A**

In 2016, our client, Fund Public Schools, attempted to pass a ballot measure in their state. In that election, voters had the opportunity to support or oppose taxing cigarettes to fund K-12 schools. Our partners want to understand what happened on election day, please explore the precinct-level data. Precinct-level results for this ballot measure are provided in precinct\_level\_election\_results.csv. Individual-level demographic features for voters registered to vote in these precincts is provided in individual\_demographics\_and\_scores.csv. Please conduct a retrospective analysis of this election.

* What factors do you think relate to support for this ballot measure? (use the precinct level aggregates found in precinct\_votes\_for\_ballot\_measure and precinct\_votes\_against\_ballot\_measure)
* What factors do you think relate to turning out to vote in this election? (precinct\_total\_ballots tells you the number of ballots cast by voters in each precinct, while precinct\_registered\_voters gives the number of registered voters in the precinct).
* What information do you think would aid our partner in voter outreach in a subsequent, similar, election?

### **Section B**

Fund Public Schools recently reached out to let us know that there is a similar ballot measure up for a vote in another state this November. They conducted a statewide poll to find out which voters may be interested in supporting this ballot measure. Data from the individual poll respondents are stored in ballot\_measure\_poll.csv.

Please explore the data provided.

* Which score(s) would you recommend that the partners use for doing voter outreach so the ballot measure will be successful?
* How should our partner use this score?
* How would you convey the value of using the targeting strategy you recommend?

## **Appendix - Data Dictionary**

**Section A:**

**Dataset 1: individual\_demographics\_and\_scores.csv**

Individual-level demographic data are provided in this file. *Each row represents data from one individual (who is registered to vote in one of the precincts from Dataset 2: precninct\_level\_election\_results.csv).*

* We’ve added some columns which you may find helpful as you explore: a (modeled) probability that the individual voter belonging to race group G, P and O (**probability\_race\_{G/P/O})**, (modeled) probabilities of having children in the voters household (**probability\_children\_in\_household),** of having household income over $100k **(probability\_income\_over\_100k),** and of not completed any post high school education **(probability\_highest\_education\_high\_school).**
* We also included the individual “model scores” from our vendor for a few model scores (these features all end in ‘\_support’). The partisan\_score gives the probability an individual supports the Democratic party, the turnout\_score gives the (modeled) likelihood the individual will vote in this type of election, and the support\_cannabis\_legalization\_score gives the likelihood of an individual supporting legislation to legalize cannabis, e.g**.** You can see more on model scores below.
* [This type of dataset would be available to us before the election took place].

**Dataset 2: precinct\_level\_election\_results.csv**

Precinct-level election data are provided in this file. *Each row represents data from one precinct!*

* You will find the name of the county where that precinct is located, the precinct number, the number of registered voters, ballots cast, votes in support of or against the ballot measure, and votes for candidate U and candidate I.
* [This type of dataset would typically be available to us a few weeks after the election took place].

**Section B: ballot\_measure\_poll.csv**

Individual-level voter data from the poll are stored in ballot\_measure\_poll.csv, the ‘support\_initiative’ column indicates which voters voiced support for the measure when asked. We provide the polled voters’ demographic data (region, county, education, SES or socioeconomic status, ethnicity, ideology, kids (whether or not the voter has kids) alongside a suite of their “model scores” from our vendor.

**Model scores**

Both **individual\_demographics\_and\_scores.csv** and **ballot\_measure\_poll.csv** datasets contain several columns which end with a ‘\_score’. There are lots of these so we won’t provide specific details on each of them. These score columns range from 0-100 and express a probability- this could be the probability of turning out in an election (E.g. a presidential turnout score indicates the likelihood a voter might vote in the upcoming presidential election), or the probability of being a supporter of a Democratic party (this one is the partisan score) or likelihood of believing in a certain issue (e.g. probability of having trust in institutions is found in the trust in institutions score). Feel free to use any of them in your analysis and do not worry too much about exact descriptions of each. Scores are always available, but election results are only available after the election.

*Reach out if you have questions or concerns!*