## Explore Colorado Voting Patterns and Demographics

## Project Topic

I will use publicly available data from the Colorado state government and the US Census Bureau to characterize Colorado counties' demographic makeup and voting behavior. The period covered will nominally be from 2010 to 2020, including the pending general election to be held on November 3, 2020. This depends, of course, on the availability of election data after the election and before the date this project is due.

Colorado has eighty counties, most of which are considered conservative, as evidenced by past elections. This is not unlike rural areas in most states. I hope to apply demographic data from the US Census to better understand the basis for this behavior.

## Possible Questions Addressed

As I get to know the data better, I will know the questions that can be answered, but in the meantime, this is a list of items to consider:

1. How do rural counties affect the outcome of national elections?
2. Have voting patterns changed over the past ten years?
3. What is the demographic makeup of the active voting population?
4. Do economic shifts affect the way people vote?

## Data Sources

This proposal includes the collection and EDA on two primary data sources:

**US Census Annual Community Survey** This survey is available for 2016. The feature set included in this dataset has NOT been finalized but would likely consist of:

* Population by race, education, income, and age categories
* Number of households with internet access
* The extent of poverty in each county as well as Supplemental Poverty Survey.

**Colorado Voter Data** These data are provided by the Colorado Secretary of State. Colorado provides a comprehensive list of reports on Colorado voting activities. A sample of these follow:

* Voter registration by status (active/non-active)
* Voter registration by party, gender, and UOCAVA Type (Uniformed And Overseas Citizens Absentee Voting Act)
* Voter registration by Senate and congressional districts
* Voter registration by party affiliation/preference. Several parties are covered:
  + American Constitution
  + Approval Voting
  + Democratic
  + Green
  + Libertarian
  + Republican
  + Unity

## Data Science Tasks

The data science tasks to be completed are Exploratory Data Analysis, Clean data set, build a variety of unsupervised cluster representation of the data as well as interpret the clustering/ All code and results will be stored in Github repository (reed54/MSDS696). This will be maintained to ensure the reproducibility of the results.

The object of the data analysis is to characterize the Colorado voters in terms of its demographics as well as urban and rural residency.

### Tools Required

1. The biggest hurdle will be to fashion the API queries to get the proper data from the US Census. I plan to use both the American Community Survey and the Decentennial Census for 2010.
2. The data from the Colorado Secretary of State is in Microsoft Excel format. This should be a lot easier to manage.

### Data

The volume of data should be relatively modest. Typically, there will be eighty samples (one for each county) and possibly up to 100 features. The election results and voter registration data will be roughly the same size.

## Milestones

The following milestones are a guide to the action items to be completed to track the progress of this project.

### Week 1

* Proposal preparation (in progress)
* Data collection as raw
* Set up GitHub repository
  + reed54/MSDS696, public (in progress)

### Week 2

* Data cleaning
* Check for outliers and missing data
* Impute missing values rather than drop samples
  + Must encode ordinal features first
* Document number of records, and features retained to model

### Week 3

* Features engineering
* Prepare data summaries: statistics, covariances
* Graphic research types that are typical for this type of data.

### Week 4

* Build models:

All of these clustering techniques will be explored to determine which (possibly more than one) allow the best insight to the data.

Cluster algorithms:

* K-Means
* Mean-Shift Clustering
* Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
* Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)
* Agglomerative Hierarchical Clustering

### Week 5

### Prepare visualizations of results

### Week 6-8

* Summarize interpretation of the clustering results.
* Finalize report and cleanup git repo.
* Finalize Presentation

## References

(In progress)