

- 주제: 유권자들의 학력 수준에 따른 후보 별 득표율 분석
- 사용할 데이터

(1) 2016 US Primary Election dataset

- Kaggle에서 제공하는 미국 대선 후보들의 2016년 예비 선거(primary)결과
- 데이터의 크기(샘플 수): 13,212개

미국 Alabama 주의 Autauga County부터 New Hampshire 주의 Sullivan County까지

- 어떤 Field가 있는지?

State, County, Party, Candidate, Vote, Vote ratio

- 어떤 기간 동안에 측정된 데이터인지?

2016년 02월 01일 Iowa 주 Kossuth County를 시작으로 공화당(Republican)은 06월 07일 California 주를 마지막으로, 민주당(Democrat)은 06월 14일 Washington D.C.를 마지막으로 측정된 데이터이다.

(2) County Facts

- 2016 US Election dataset에 포함되는 각 주의 인구 유형 통계
- 데이터의 크기(샘플 수): 3,195개

- 어떤 Field가 있는지?

- Area name, EDU635213, EDU685213

EDU635213: High school graduate or higher, percent of persons age 25+, 2009-2013

EDU685213: Bachelor's degree or higher, percent of persons age 25+, 2009-2013

- 분석 방법

학력 수준이 같고 지지하는 후보가 동일하면 득표 수를 더하는 방법으로 Dataset을 정리한다.

학력 수준 별 지지하는 후보의 득표율을 알아보기 위해 해당 학력 수준 집단에서 각 후보의 득표 수 Data를 입력하여 학력-후보 간의 득표 수 Table을 만든다.

이 Table을 통해 각 후보를 지지하는 집단의 학력 수준 분포를 알아본다.

이 Table을 통해 특정 학력 집단에서의 각 후보 별 지지 비율을 알아본다.

이 과정을 통해 학력 수준과 후보 별 득표율의 관계를 알아본다.

- 예상 결과

공화당 지지자와 민주당 지지자 간 학력 수준의 차이는 예상해 볼 만한 기준이 없으나 도널드 트럼프 후보의 당시 중하층 백인 유권자들을 대상으로 펼쳤던 선거 전략 중 하나인 아웃사이더 캠페인의 영향으로 해당 후보를 지지하는 집단의 학력 수준이 다른 후보의 경우보다 비교적 낮을 것으로 예상된다.