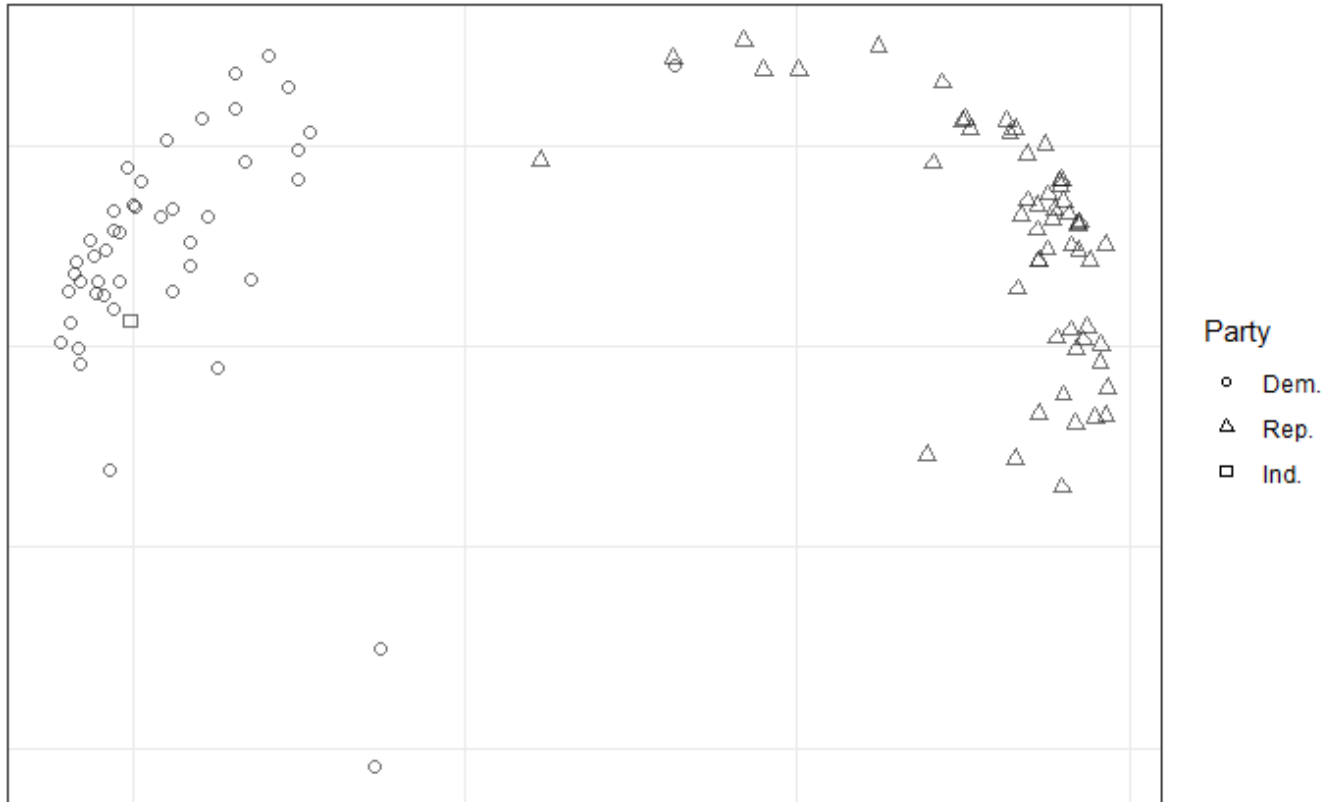


배경

책 <Machine Learning for Hackers>의 9장을 참고해, MDS(Multi-Dimensional Scaling)로 미 상원 의원들의 유사성을 시각적으로 탐색하는 과정을 학습해보았고, 101회부터 110회의 국회 중 약 5회에 서 아래와 같이 양극화된 모습을 볼 수 있었다.

Roll Call Vote MDS Clustering for 110th U.S. Senate



110회

그래프들을 보며 우리나라의 경우는 어떨까 하는 생각이 들어, MDS 프로세스를 복습해볼 겸 이번 분석을 진행하게 되었다.

데이터

우리나라 국회 표결 데이터는 [열린국회정보 정보공개포털](https://yroh.h.tistory.com/m/210)에서 2021년 1월 건만 내려받아 사용하였다.

툴

Excel과 R Studio를 사용하였다.

패키지

- reshape
- dplyr
- ggplot2

프로세스 요약

의원-의안번호 행렬 생성

행렬곱(원행렬*전치행렬)

유클리드 거리 행렬

MDS 적용

시각화

상세 프로세스

엑셀에서 분석에 필요한 의안번호, 의원, 정당, 표결결과를 제외한 나머지 열들을 모두 제거

의안번호의 타입을 factor로 변경

무소속을 포함한 8개의 정당 중 한 명의 당원만 존재하는 '시대전환'과 '기본소득당'을 '무소속'으로 변경

* 그래프에 값들을 당으로 구분하여 나타낼 시, 구분할 수 있는 수가 6개가 최대이므로 위와 같이 변경

정당의 타입을 factor로 변경

표결결과의 값을 찬성은 1로, 반대는 -1로, 기권은 0으로 변경한 뒤, 표결결과의 타입을 numeric으로 변경

좌측 데이터의 모양을 cast 함수를 이용하여, 우측과 같이 의원-의안번호 행렬로 변환

의안번호	의원	정당	표결결과
1 2104473	강대식	국민의힘	1
2 2104473	강득구	더불어민주당	1
3 2104473	강민국	국민의힘	1
4 2104473	강병원	더불어민주당	1
5 2104473	강선우	더불어민주당	1
6 2104473	강훈식	더불어민주당	1


```
> head(data.matrix)
```

	2101114	2101748	2103134	2103623	2103714	2103805	2103820	2104203	2104473	2104557	2107249	2107250	2107251	2107252
강대식	1	1	1	1	1	1	1	1	1	1	1	1	1	1
강득구	1	1	1	1	1	1	1	1	1	1	1	1	1	1
강민국	1	1	1	1	1	1	1	1	1	1	1	1	1	1
강민정	1	1	1	1	1	1	1	1	1	1	1	1	1	1
강병원	1	1	1	1	1	1	1	1	1	1	1	1	1	1
강선우	1	1	1	1	1	1	1	1	1	1	1	1	1	1

전치행렬을 이용한 행렬곱 -> dist함수 거리 산출 -> cmdscale함수 좌표 추출

```
> head(mp.data.matrix.dist.mds)
```

	[,1]	[,2]
강대식	-17.00929	-0.2138262
강득구	-17.00929	-0.2138262
강민국	-17.00929	-0.2138262
강민정	-17.00929	-0.2138262
강병원	-17.00929	-0.2138262
강선우	-17.00929	-0.2138262

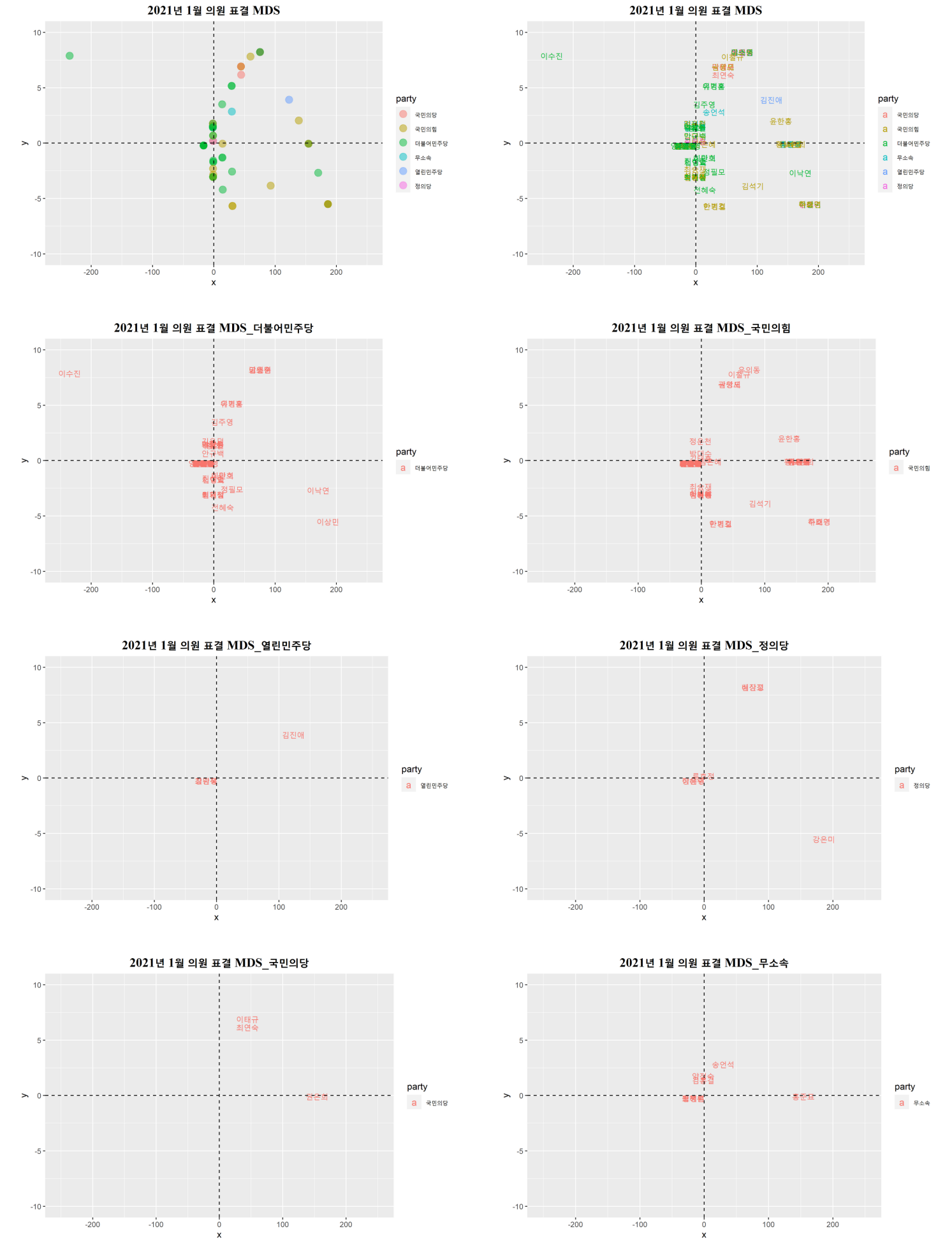
시각화 편의를 위한 데이터프레임 생성

```
> head(rollcall.mds)
```

	x	y	name	party
1	-17.00929	-0.2138262	강대식	국민의힘
2	-17.00929	-0.2138262	강득구	더불어민주당
3	-17.00929	-0.2138262	강민국	국민의힘
4	-17.00929	-0.2138262	강민정	열린민주당
5	-17.00929	-0.2138262	강병원	더불어민주당
6	-17.00929	-0.2138262	강선우	더불어민주당

그래프 시각화

결과



결론

이번 분석에 적용된 데이터는 여러 문제를 포함하고 있으며, 위의 도출된 MDS 그래프는 유의한 인사이트를 도출하기엔 부적합한 것으로 보인다.

1. 다른 의원들과의 유사성이 상당히 적어 보이는 이수진 의원의 데이터를 들여다보니, 더불어 민주당 내에 동명의 이수진 의원이 한 명 더 존재한다는 것을 알 수 있었다(이수진 의원뿐만 아니라, 다른 이름도 존재할 것으로 보인다).
2. 의안 별로 표결에 참여한 의원의 수가 다르다.

우선 결과에 가장 관련된 문제는 위 두 가지라고 생각한다. 데이터셋을 가공하여 위의 두 문제를 해결한 뒤, 동일 프로세스를 적용하게 되면 다른 모습의 그래프가 도출될 것으로 보인다. 하지만 이번 분석에선 MDS에 적합한 데이터의 형태와 프로세스를 직접 경험해봤다는 것에 만족하고, 추후 다른 주제를 통해 MDS를 적용해볼까 한다.