
경영경제 데이터 분석

소프트웨어


기말고사 보고서



D e p a r t m e n t : 경 영 학


S t u d e n t I D : 2 0 1 4 3 4 0 8


N A M E : 방 준 석

	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	

목차

1	서론	4
1.1	주제	4
1.2	연구 배경	4
1.3	체스에 대한 간략한 소개	4
1.4	연구 목적	6
2	데이터 소개.....	7
2.1	출처와 해당 기관.....	7
2.2	데이터 살펴보기.....	7
2.3	Column 소개.....	8
3	분석	9
3.1	가장 많이 사용 된 첫 번 째 수는 무엇일까?.....	9
3.2	가장 많이 사용되는 오프닝	12
3.3	‘백’ 이 승리 많이 하는 오프닝	13
3.4	‘흑’ 이 승리 많이 하는 오프닝	14
4	논의	15
4.1	데이터 분석을 하면서 느낀 점.....	15
4.2	데이터 분석의 한계점	15
5	참고자료	16

	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	

	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	

1 서론

1.1 주제

이번 데이터 분석 프로젝트의 대상은 많은 사람들이 조금은 생소하게 느낄 수 있는 **“Chess”** 이다. 우리나라에서는 바둑, 장기에 비해 인기가 현저히 떨어지지만, 미국, 영국, 인도, 러시아를 비롯한 대부분의 서양 국가들에서는 오랫동안 큰 인기와 전통을 이어 가고있다.

1.2 연구 배경


사실 보고서의 주제로 체스를 정한 이유는 그다지 복잡하지 않다. 본인은 체스를 어릴 때부터 지금까지 취미로 꾸준히 공부 해왔고, 해외에서 대회 경험 및 수상 경력도 어느정도 있을 만큼 관심이 많다.

현시대의 프로 체스 선수들(Grandmaster, International Master) 은 데이터분석 및 인공지능 없이 체스 대회를 준비 할 수 없을 만큼 체스는 데이터분석과 밀접한 관련이 있다는 점을 알아둘 필요가 있다. 체스 경기 데이터를 통해서 선수들은 수 많은 통찰을 얻을 수 있고, 인공지능을 활용해서 자신의 실수들을 파악 할 수 있다.

따라서 이번 프로젝트를 계기로 데이터분석을 활용하여 체스 데이터로부터 어떠한 분석들을 할 수 있을지 알아 보고싶었다.

1.3 체스에 대한 간략한 소개

이번 프로젝트의 연구 목적 및 데이터 분석을 본격적으로 살펴보기에 앞서, 보고서 이해를 돕기 위해 체스에 대한 간단한 소개를 하겠다.

	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	




하나의 체스 경기는 크게 오프닝, 미드게임, 엔드게임으로 구성 되어있다.

오프닝은 경기의 시작 단계로, 각 플레이어들은 이 단계부터 각자 준비했거나 자신 있는 전략들을 사용하기 시작한다.

미드게임은 경기의 중간 단계로, 가장 활발하게 서로를 공격하고 수비하기 시작하는 단계라고 볼 수 있다.

엔드게임은 경기를 마무리 짓는 단계로, 몇 개 남지 않은 체스 piece 를 활용하여 승패를 결정하기 위해 싸우는 마지막 단계다.

* “Avengers: End Game”의 제목에서 사용하는 엔드게임 또한 체스에서 유래

	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	


1.4 연구 목적



<Image 1.1>

체스 한 게임이 시작하고 두 선수가 각자 수를 두면 발생 할 수 있는 체스 게임의 경우의 수는 400 개가 된다. 그 다음에 또 양쪽 선수가 수를 두면 197, 742 개, 거기서 한 번 더 두면 12,100,000 개의 체스 게임이 가능해진다. 이론적으로 한 체스 게임에서 가능한 경우의 수는 우주에 있는 분자 수보다 많다고 할 만큼, 시작하자마자 정말 다양한 오프닝을 미리 준비 해야된다.

체스에는 총 1327 개의 오프닝 전략이 존재하고 그 중에서 유명한 것들은 책 한권 씩 쓸 수 있는 분량의 이론이 있다. 이번 데이터분석을 통해 어떤 오프닝들이 가장 인기가 많은지, 그리고 어떤 오프닝을 쓰는 것이 가장 승률을 높이는지 알아보려고 한다.

	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	

2 데이터 소개

2.1 출처와 해당 기관

본 보고서의 데이터는 Kaggle 에서 가져온 데이터로, 가장 인기 많은 체스 플랫폼인 Lichess 의 API 에서 추출 된 데이터이다.

Lichess 데이터베이스에는 총 1,698,693,105 개의 체스 게임 데이터가 기록되어 있고, 각 게임에서 사용된 모든 오프닝 뿐만 아니라 모든 수에 대한 정보가 전부 기록 되어있다.

*각 기관의 링크는 마지막 참고 자료 섹션에 기재 되어있다

2.2 데이터 살펴보기

본 데이터파일을 불러와서 데이터를 살펴보자.

```
> dim(data)
[1] 20058 16
```

<Image 2.1>


총 16 개의 column 과 20058 개의 row 로 구성되어 있는것을 dim() 으로 확인할 수 있다.

id	rated	created_at	last_move_at	turns	victory_status	winner	increment_code	white_id	white_rating	black_id	black_rating	moves
1	TZHLUE	FALSE	1.50421e+12	1.50421e+12	13	outoftime	white	15+2	bourgris	1500	a-00	1191 d4 d5 c4 c6 cxd5 e6 dxe6 fxe6 Nf3 Bb4+ Nc3 Ba5 Bf4
2	IINXwaE	TRUE	1.50413e+12	1.50413e+12	16	resign	black	5+10	a-00	1322	skinnerua	1261 d4 Nc6 e4 e5 f4 f6 dxe5 fxe5 Nxe5 Qd4 Nc6 Qe.
3	mICvQh	TRUE	1.50413e+12	1.50413e+12	61	mate	white	5+10	ischia	1496	a-00	1500 e4 e5 d3 d6 Be3 c6 Be2 b5 Nd2 a5 a4 c5 axb5 Nc6 b.
4	kWKvrqYL	TRUE	1.50411e+12	1.50411e+12	61	mate	white	20+0	daniamurashov	1439	adivanov2009	1454 d4 d5 Nf3 Bf5 Nc3 Nf6 Bf4 Ng4 e3 Nc6 Be2 Qd7 O-O.
5	9xoIAUZ	TRUE	1.50403e+12	1.50403e+12	95	mate	white	30+3	nik221107	1523	adivanov2009	1469 e4 e5 Nf3 d6 d4 Nc6 d5 Nb4 a3 Na6 Nc3 Be7 b4 Nf6 .
6	MooDV9wi	FALSE	1.50424e+12	1.50424e+12	5	draw	draw	10+0	trtkon17	1250	franklin14532	1002 e4 c5 Nf3 Qx5 a3

<Image 2.2>

opening_eco	opening_name	opening_ply
D10	Slav Defense: Exchange Variation	5
B00	Nimzowitsch Defense: Kennedy Variation	4
C20	King's Pawn Game: Leonardis Variation	3
D02	Queen's Pawn Game: Zukertort Variation	3
C41	Philidor Defense	5
B27	Sicilian Defense: MongOOSE Variation	4


<Image 2.3>

	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	

2.3 Column 소개

- *Game ID*: 게임 고유 id
- *Rated (T/F)*: 체스에는 선수들이 Rating 이 각자 있다. 게임에서 승리하면 Rating 은 올라가고 패하면 내려간다. 게임 결과에 따라서 Rating 에 변화가 있는 게임도 있고 없는 게임도 있기 때문에, 이 칼럼은 그 유무를 표시한다.
- *Start Time*: 게임이 시작한 시간
- *End Time*: 게임이 끝난 시간
- *Number of Turns*: 총 발생한 체스 수
- *Game Status*: 게임이 어떻게 마무리 되었는지. (시간 초과, 기권, 체크메이트, 무승부 등이 있다)
- *Winner*: 승자
- *Time Increment*: 체스 게임 시간 포맷
- *White Player ID*: 백색 플레이어 id
- *White Player Rating*: 백색 플레이어 레이팅
- *Black Player ID*: 흑색 플레이어 id
- *Black Player Rating*: 흑색 플레이어 레이팅
- *All Moves in Standard Chess Notation*: 모든 체스 수
- *Opening Eco*: 오프닝 고유 코드
- *Opening Name*: 오프닝 이름
- *Opening Ply*: 해당 오프닝과 관련된 수

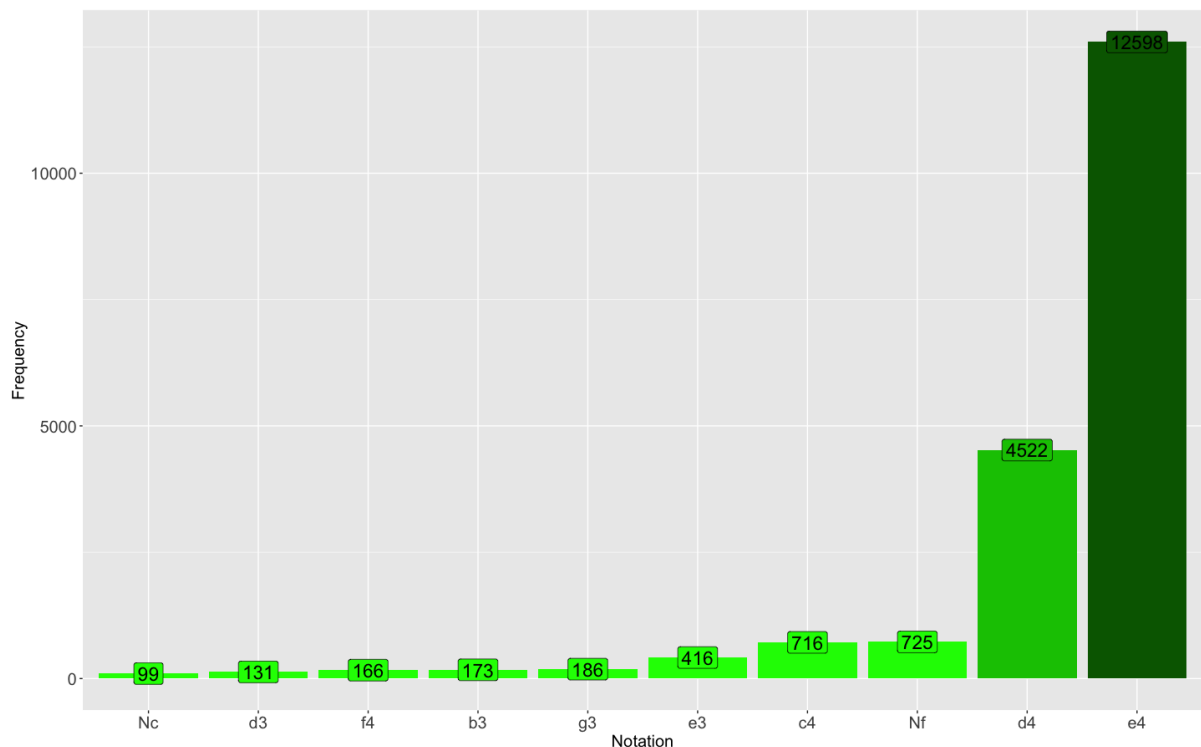
이 데이터들을 사용해서 어떤 오프닝이 가장 효과적인지 알아볼 차례다.

	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	

3 분석

3.1 가장 많이 사용 된 첫 번째 수는 무엇일까?

가장 많이 사용되는 오프닝을 분석 하기에 앞서, 가장 많이 사용 되는 “첫 수”가 무엇인지 알아 보는것도 의미가 있다. 백색 플레이어가 두는 첫 수에 따라서 어떤 오프닝으로 게임이 진행 될지가 결정 되기 때문이다




<Figure 3.1>

위 그래프의 y 축은 빈도 수, x 축은 Notation 이다.

*Notation 은 체스에서 사용되는 표기 법이다

위 데이터에서 도출 할 수 있는 결과로는, 백색 플레이어는 **e4** 와 **d4** 를 다른 수에 비해 압도적으로 많이 사용한다는 것이다. 이것은 체스를 어느 정도 하는

	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	

사람들이 충분히 체스를 하면서 확인 할 수 있는 사실이다. 체스에서 가장 인기 많은 오프닝이 보통 e4 와 d4 로 시작하기 때문이다.




<Image 3.1>

위 사진에서 백색 플레이어가 둔 수가 바로 **e4** 이다.

여기서 흑색 플레이어가 자주 사용하는 수는 e5, c5 (화살표가 표시하는 수) 가 있고, 여기서 수 많은 오프닝 전략 이 시작된다.

E4 로부터 시작되는 인기 많고 유명한 오프닝의 종류로는 Ruy Lopez, Sicilian Defence(화살표가 표시하는 수에서 시작한다), Italian Game, Latvian Gambit, Petrov's Defence, Scotch Game, 등등 수 없이 많다. 각 오프닝 이론 내에서도 많은 종류의 전략들이 있는 데 이것을 보통 “line” 이라고 부른다. 체스 선수들은 이 모든 line 의 20~30 수 는 기본적으로 전부 외우고 있다.

	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	




<Image 3.2>

위 사진에서 나온 수가 두 번 째로 많이 사용 되는 **d4** 이다. d4 도 e4 와 마찬가지로 높은 수준의 체스에서 정말 많이 나타나는 수로, 수 백 가지의 오프닝의 시작점이다.



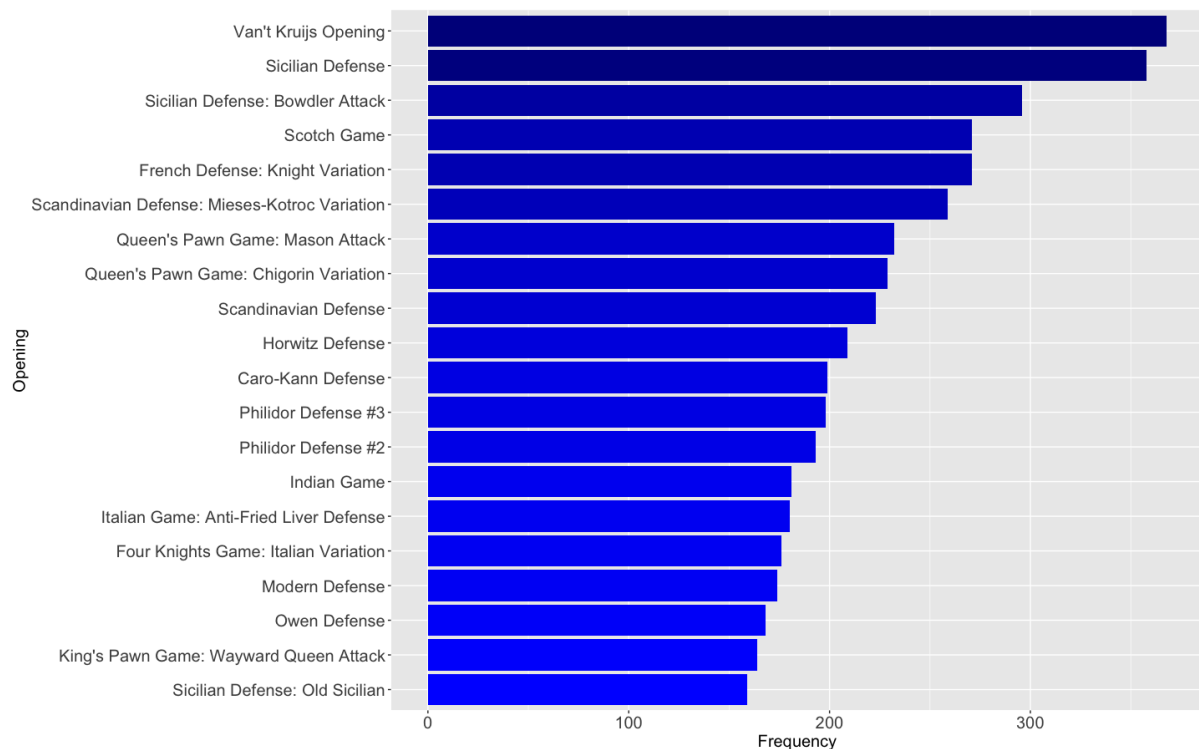
<Image 3.3>

 중앙대학교	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	

위 사진은 요즘 넷플릭스에서 인기 많은 체스 드라마의 제목으로 유명한 Queen's Gambit 의 시작 상태이다. 이 오프닝이 바로 d4 로 시작 되는 것이다.


3.2 가장 많이 사용되는 오프닝

다음 분석 결과로는 Lichess 에서 가장 많이 사용되는 오프닝을 20 위 까지 그래프로 표현 했다.



<Figure 3.2>

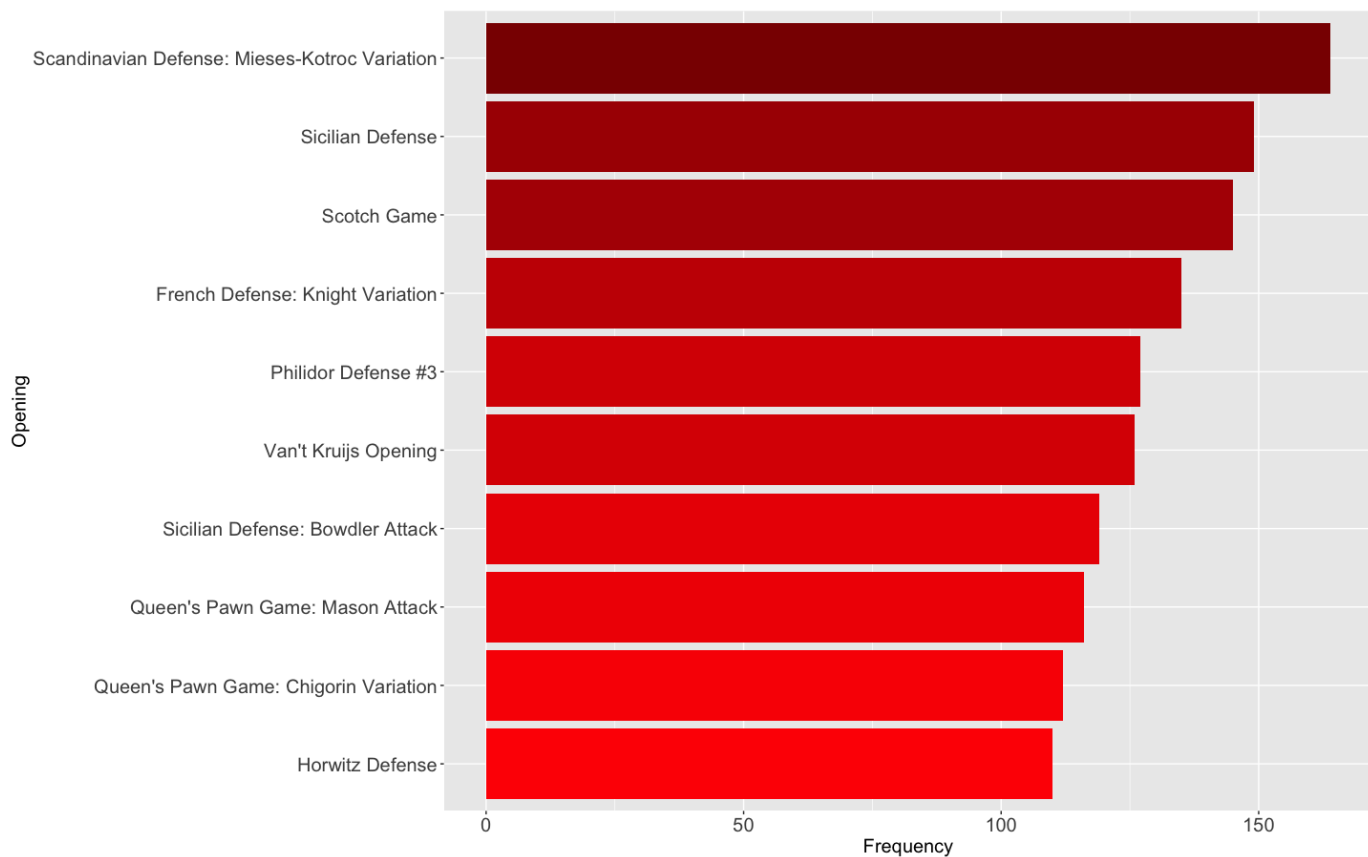
Lichess 에서 가장 많이 사용되는 오프닝으로는 Van't Kruijs Opening, 그 다음으로는 Sicilian Defense 가 자리 잡고 있다. 물론 두 개의 오프닝 모두 효과적이고 승률도 높은 오프닝이지만, 이 그래프는 승률과는 상관없이 “재미”, “공부” 와 같은 목적으로 사람들이 많이 사용 하는 거 일 수도 있다. 그렇기 때문에 다음 분석으로 ‘백’ 이 가장 많이 이기는 오프닝, 그리고 ‘흑’이 가장 많이 이기는 오프닝을 분석 해보겠다.

	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	

3.3 ‘백’ 이 승리 많이 하는 오프닝


이제부터 진짜 승률과 상관 있는 데이터 분석이다. 데이터 분석을 해본 결과, ‘백색’ 플레이어가 이길 수 있는 가장 좋은 오프닝은 Scandinavian Defense: Mieses-Kotroc Variation 이 되겠다.

실제로 수준 높은 대회에서 많이 사용되는 체스 오프닝으로, 백색이 매우 공격적으로 게임을 임하는 것이 특징이기 때문에, 승리를 위해 선택되는 경우가 많다.



<Figure 3.3>

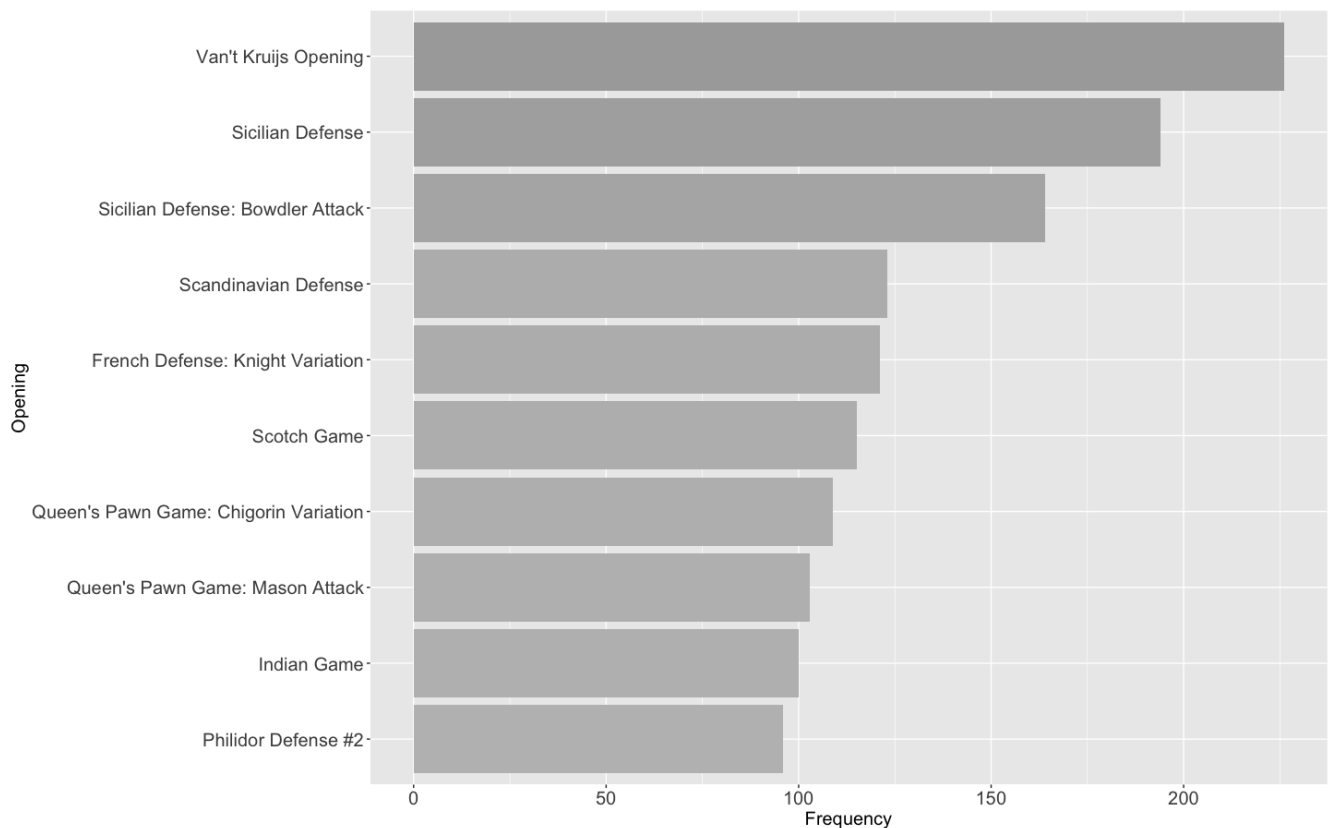
높은 레벨의 체스에서는 (세계적인 선수들의 대회) 공격적인 오프닝을 선택해서 무조건 승점을 따내려고 하는 선수가 있는 방면(대체적으로 ‘백’), 무승부를 노리고 수비를 선택하는 선수도 많다. 공격적으로 게임 운영을 할 것인지 수비적으로 할 것인지 결정하는 것이 ‘백’ 선수의 몫이 되는 것이다.

	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	

이러한 점을 고려했을 때, 대부분 공격적인 성향을 갖고있는 오프닝으로 데이터 분석 결과가 나왔다는 것이 흥미롭다.


3.4 ‘흑’ 이 승리 많이 하는 오프닝

마지막 데이터 분석 결과로 ‘흑’ 색의 선수가 승리 많이 하는 순위를 그래프로 표현했다.



<Figure 3.4>

‘흑’의 승률이 가장 높은 오프닝으로 ‘Van’t Kruijs Opening’ 과 ‘Sicilian Opening’ 인 것을 확인 할 수 있다. 여기서 흥미로운 점은, Van’t Kruijs Opening 은 세계적인 수준에서 자주 사용되는 Opening 은 아니라는 점이다. Ruy Lopez, Sicilian

	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	

Defense, Queen's Gambit 과 같은 유명한 오프닝과는 다르게 아는 사람이 많지 않다. 또 한가지 흥미로운 점은, 이 오프닝은 앞선 분석 결과에서 가장 널리 사용되는 체스 오프닝으로도 나왔다는 것이다. 왜 유명하지 않은 오프닝이 가장 많이 사용 되고, ‘흑’이 승리하기 가장 좋은 오프닝으로 나온 것인지 생각 해볼 필요가 있다.

2 위로 나온 ‘Sicilian Defense’ 는 체스에서 가장 유명한 오프닝이기도 하고, ‘흑’ 을 두는 플레이어가 사용 할 수 있는 가장 공격적인 오프닝으로 잘 알려져 있다. 넷플릭스 드라마 ‘퀸스 갬빗’의 주인공이 가장 많이 사용하는 오프닝이기도 하다.


4 논의

4.1 데이터 분석을 하면서 느낀 점

이번 데이터 분석을 하면서 체스 오프닝에 대한 다양한 통찰이 생겼다. 체스 게임에서 이기기 위해 어떠한 오프닝을 사용하면 좋은지, 그리고 반대로 패하지 않기 위해서는 어떤 오프닝에 대한 준비가 되어있어야 하는지 확인하기 좋은 데이터 분석이라고 생각한다.

4.2 데이터 분석의 한계점

이 데이터 분석의 한계점으로는 다소 부족한 데이터의 양이 될 수 있다고 생각한다. 데이터의 질이 상당히 좋고 활용 할 수 있는 방법이 여러가지라서 좋았지만, 전체적인 양이 조금은 부족 했다고 느꼈다. 오프닝의 종류만 1000 개가 넘는데 전체 데이터의 양이 그것의 두배 밖에 안됐다. 실제로 Lichess 에서 진행된 모든 체스 게임을 가지고 데이터 분석을 할 수 있었다면, 조금은 다른 결과가 나왔을지도 모르겠다는 생각도 드는 부분이다.

	경영경제 데이터분석 소프트웨어			방준석
	Category	Version	Reporting date	
	경영경제 데이터분석 소프트웨어	3.0	2020. 12. 17	

5 참고자료

참고 자료

<https://www.kaggle.com/datasnaek/chess> (데이터 출처)

<https://lichess.org/> (체스 게임 이미지 캡처)

프로그래밍 참고 자료

<https://www.datacamp.com/community/tutorials/pipe-r-tutorial>

<https://ggplot2.tidyverse.org/reference/>

<https://dplyr.tidyverse.org/reference/mutate.html>

<https://dplyr.tidyverse.org/reference/arrange.html>

<http://zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2-cheatsheet-3/>