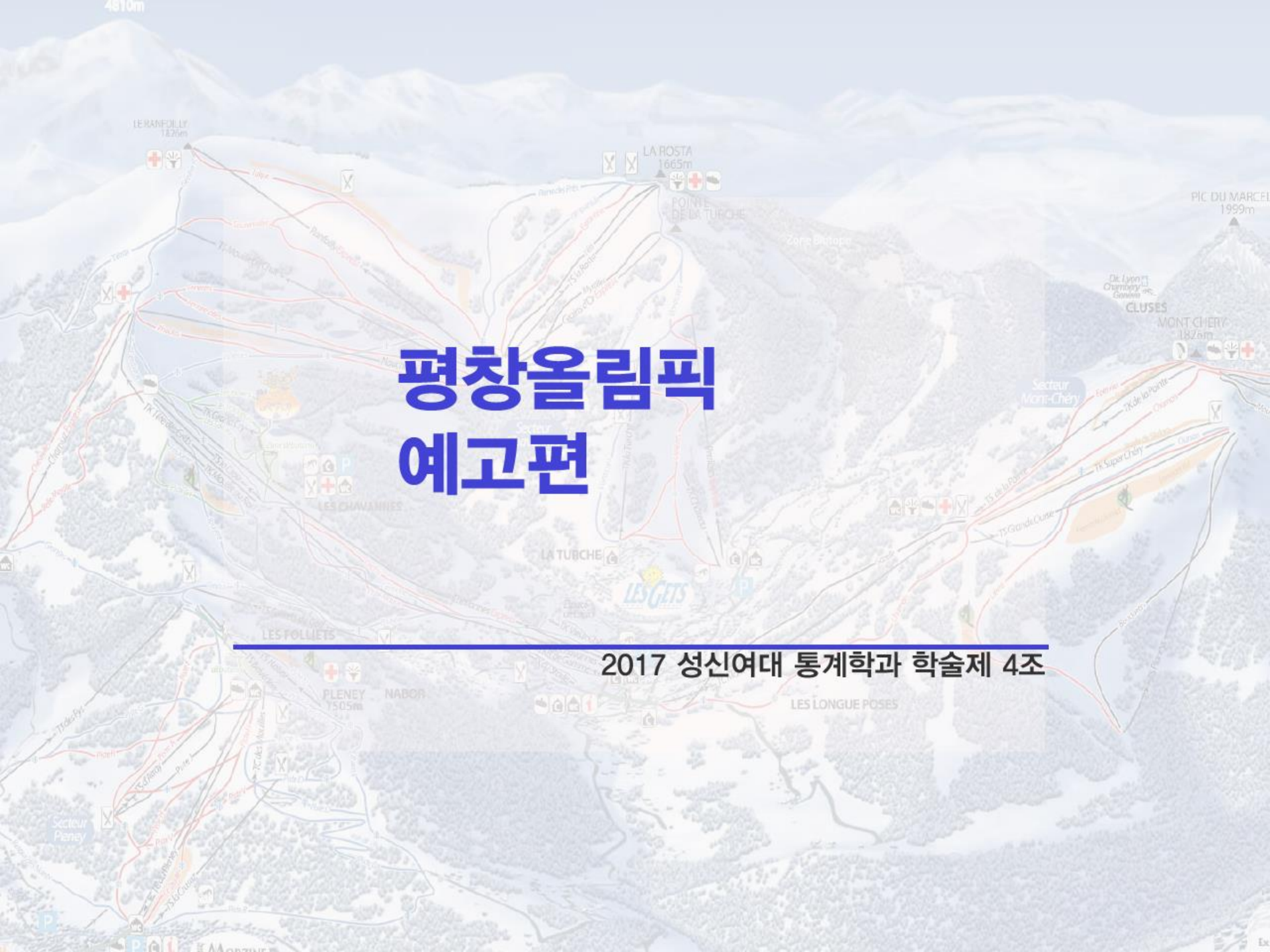


평창올림픽 예고편

2017 성신여대 통계학과 학술제 4조

목차

주제 선정 과정

통계적 모델링

후기

팀 소개

데이터 수집

결론

1 팀 소개

2 주제 선정

3 데이터 수집

4 통계적 모델링

5 결론

6 후기

조장	김유경			
15학번	김지원	김태연	최다영	최지수
16학번	김진현	박수인	성수연	오아연
	전은지	홍예지		
17학번	김경희	김주선	박규빈	박소은
	정지우	정세영		

팀 소개

데이터 수집

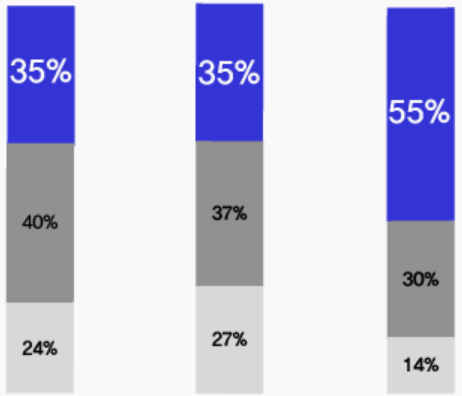
결론

2. 주제 선정 평창 동계올림픽

2. 주제 선정 국민 인식 조사

문화체육관광부 발표 '평창올림픽 국민 여론 조사'

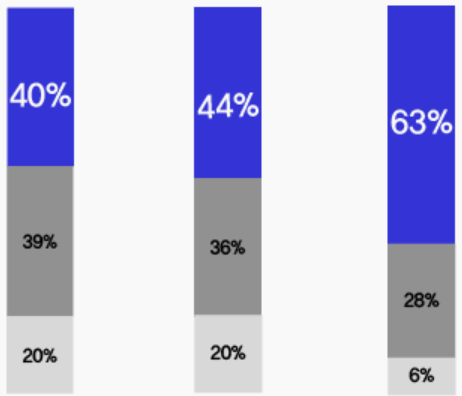
▶ 1차 여론조사



기간 :17.03.24~17.03.25

그렇다 보통이다 아니다

▶ 2차 여론조사



기간 : 17.05.26~17.05.27

그렇다 보통이다 아니다

▶ 조사 방법 무작위 선정 유무선 전화 방식

▶ 대상 및 규모 전국 15세 ~ 79세, 일반국민 1000명

프로젝트 목표

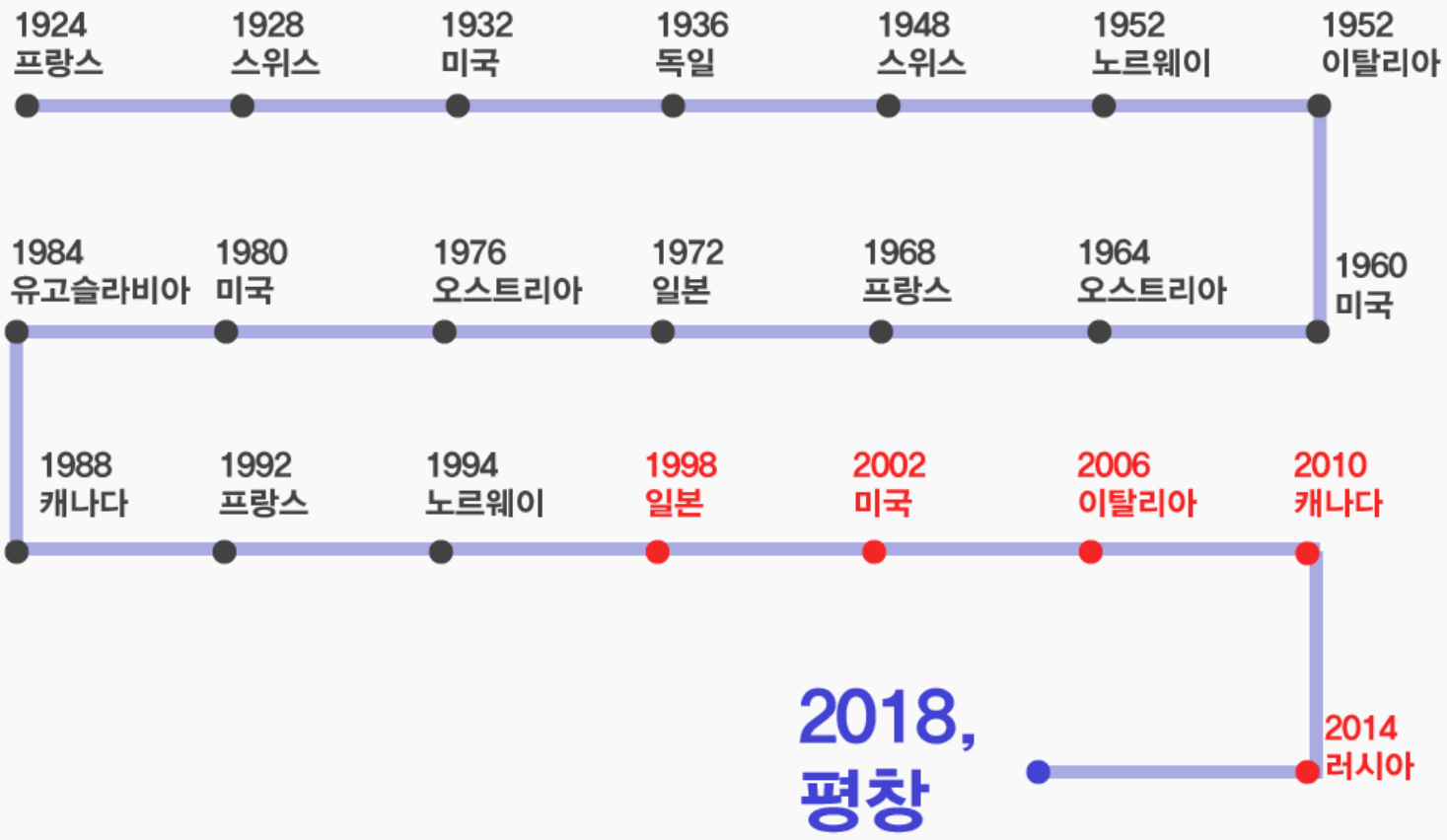
평창동계올림픽 메달 수 예측

▶ 통계모형을 이용해보자!


주제 선정
통계적 모델링
후기
1
2
3
4
5
6
팀 소개
데이터 수집
결론

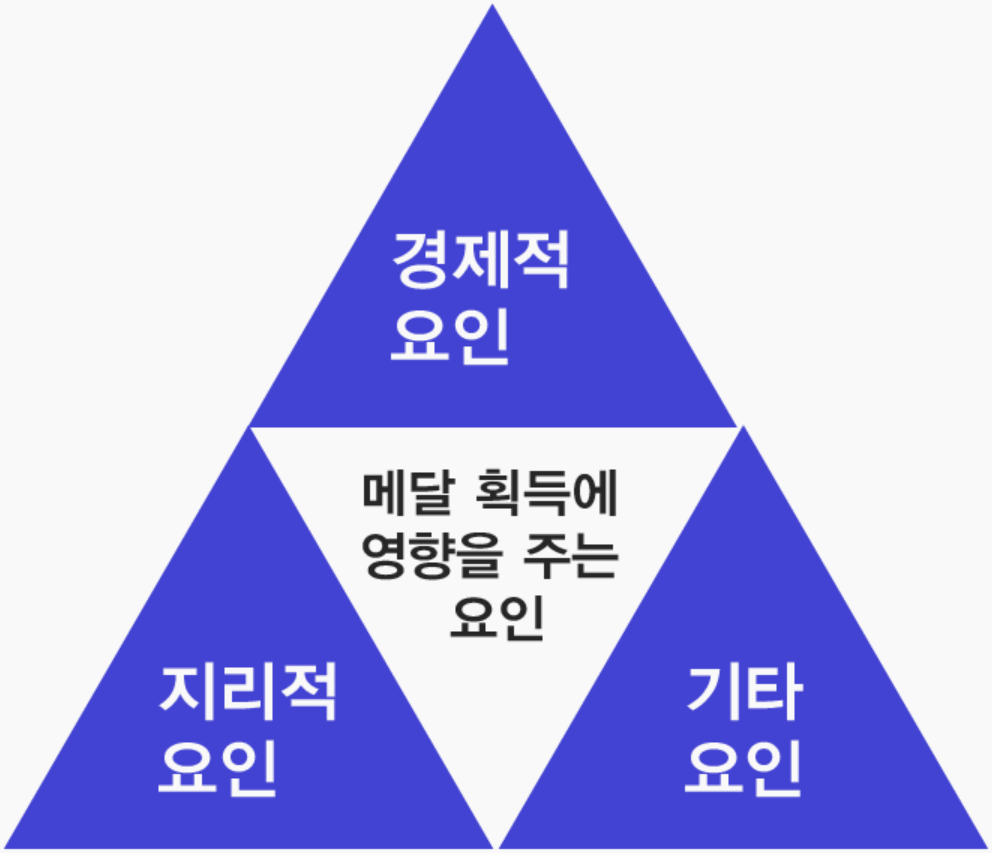
3. 데이터 수집

기간 설정


1924
프랑스
1928
스위스
1932
미국
1936
독일
1948
스위스
1952
노르웨이
1952
이탈리아
1960
미국
1964
오스트리아
1968
프랑스
1972
일본
1976
오스트리아
1980
미국
1984
유고슬라비아
1988
캐나다
1992
프랑스
1994
노르웨이
1998
일본
2002
미국
2006
이탈리아
2010
캐나다
2014
러시아
2018,
평창

3. 데이터 수집

변수 선정



3. 데이터 수집 변수 선정

경제적 요인

GDP 1인당 GDP

GNI 1인당 GNI

수출량

▶ 출처 : 세계 은행

THE WORLD BANK
IBRD • IDA

3. 데이터 수집 변수 선정

지리적 요인

▶ 출처 : CIA

국가 면적

위도

3. 데이터 수집

변수 선정

기타 요인

참가선수 수

평균 경기 실적

주최국 효과

기대 수명

▶ 출처 : 세계 은행 & Kaggle

THE WORLD BANK

kaggle

3. 데이터 수집 수집한 데이터 예시 자료

1	메달합계	국가코드	국가명	년도	참가선수 수	국가 면적	위도의 절대값	GDP	1인당 GDP	인구 수	직전 경기 실적	수출량	GNI	1인당 GNI	기대수명	주최국
2	29	GER	독일	1998	125	357022	51.165691	2243225520	27.34067289	82047195	24	593408914	2.30E+12	27990	77.476	0
3	25	NOR	노르웨이	1998	76	323802	60.472024	154165219.8	34.78877856	4431464	26	56898782	1.60E+11	36110	78.329	0
4	18	RUS	러시아	1998	122	17098242	61.52401	270953117	1.834846938	147670692	23	84595569	3.15E+11	2130	69.73	0
5	15	CAN	캐나다	1998	144	9984670	56.130366	631813279.4	20.88783947	30247900	13	253008426	6.28E+11	20760	78.662	0
6	13	USA	미국	1998	186	9826675	37.09024	9089168000	32.94919776	275854000	13	952981000	8.87E+12	32150	76.58	0
7	11	NED	네덜란드	1998	22	41543	52.132633	432476116.4	27.53360679	15707209	4	258337036	4.39E+11	27930	77.883	0
8	10	JPN	일본	1998	156	377915	36.204824	4032509761	31.9027671	126400000	5	424400693	4.35E+12	34450	80.501	1
9	17	AUT	오스트리아	1998	96	83871	47.516231	217683626.1	27.28963071	7976789	9	83704902	2.23E+11	27920	77.671	0
10	6	KOR	대한민국	1998	37	99720	35.907757	376481975.7	8.133731246	46286503	6	151153456	4.66E+11	10070	74.68	0
⋮																
391	0	TUR	터키	2014	6	783562	38.963745	798781754.1	10.30369871	77523788	0	221998757	9.76E+11	12590	75.152	0
392	0	TGO	토고	2014	2	56785	8.619543	4482880.424	0.630046061	7115163	0	1994812	3842188020	540	59.576	0
393	0	TON	통가	2014	1	747	21.178986	443475.1421	4.200132045	105586	0	82041	461410820	4370	72.701	0
394	0	PRY	파라과이	2014	1	406752	23.442503	30881166.85	4.712870205	6552518	0	13954911	28765554020	4390	72.932	0
395	0	PAK	파키스탄	2014	1	796095	30.375321	244360888.8	1.32055355	185044286	0	29916086	2.57E+11	1390	66.149	0
396	0	PER	페루	2014	3	1285216	9.189967	201021342.5	6.490181189	30973148	0	45392392	1.96E+11	6330	74.485	0
397	0	PRT	포르투갈	2014	2	92090	39.399872	229629822.1	22.07753613	10401062	0	92022714	2.21E+11	21260	81.122	0
398	0	PHL	필리핀	2014	1	300000	12.879721	284834199.3	2.873088189	99138690	0	82281140	3.44E+11	3470	68.87	0
399	0	HUN	헝가리	2014	16	93028	47.162494	139294565.1	14.11797668	9866468	0	123491978	1.33E+11	13460	75.763	0
400	0	HKG	홍콩	2014	1	1104	22.396428	291228511.4	40.21548965	7241700	0	621071949	2.91E+11	40170	83.98	0

주제 선정

통계적 모델링

후기

팀 소개

데이터 수집

결론

4. 통계적 모델링

EDA

각 변수의 결측률

참가선수 수	국가 면적	위도	수출량	GDP	1인당 GDP
0.75%	0%	0.5%	10.28%	6.02%	6.02%

GNI	1인당 GNI	인구 수	평균 경기 실적	주최국 효과	기대수명
8.77%	6.02%	2.01%	0%	0%	8.77%

결측 이유
- ▶ 현재 사라진 국가의 경우
- ▶ 공식 사이트에 누락

→ 제거!

4. 통계적 모델링

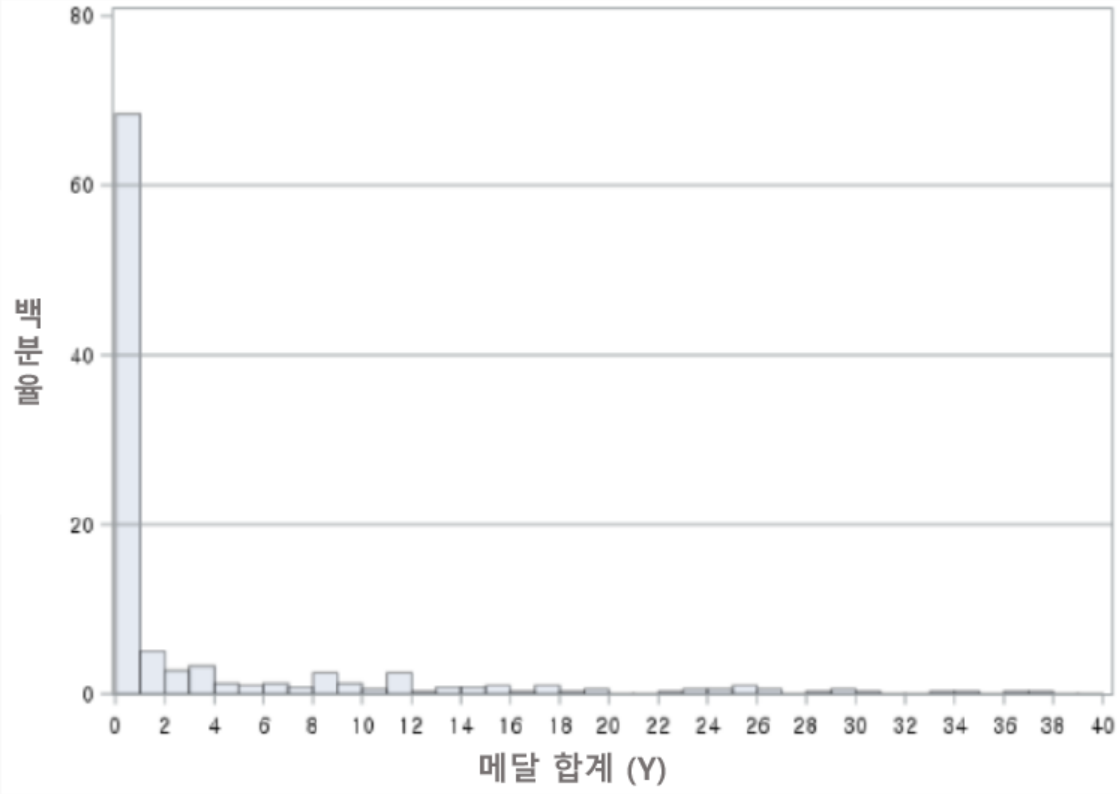
EDA

Y의 분포

▶ Y의 평균과 분산

평균 : 3.12 분산 : 46.72

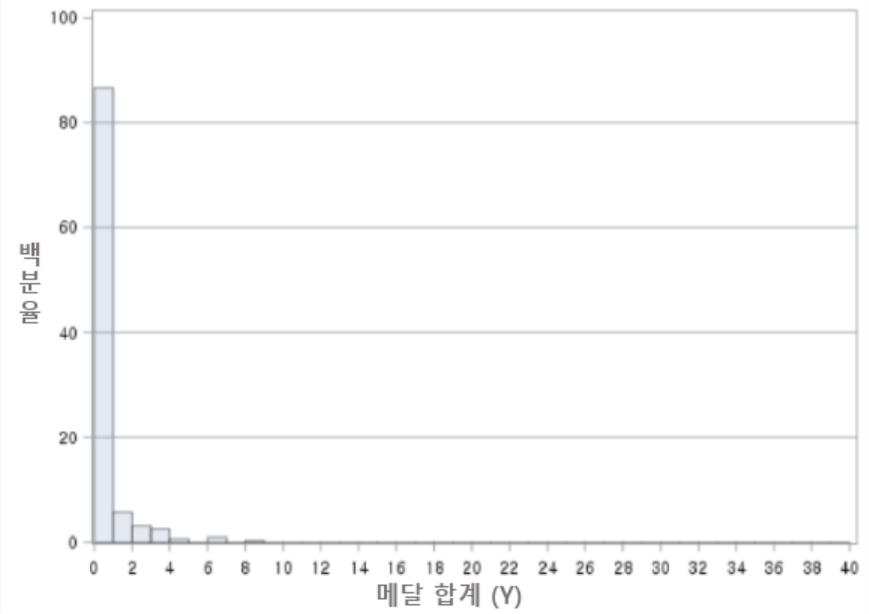
▶ 금:은:동 = 1:1:1

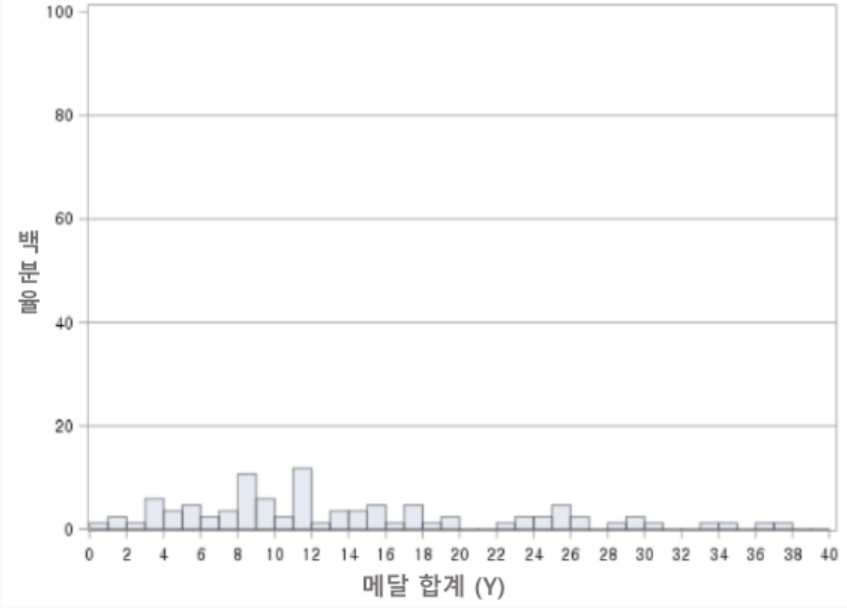


4. 통계적 모델링

EDA

평균 경기 실적에 대한 Y의 조건부 분포





▶ 경기 실적 < 3

▶ 경기 실적 ≥ 3

4. 통계적 모델링 예측 모델

▶ 데이터 분할

Train : Test = 6 :4

4. 통계적 모델링

영과잉 모형

$$f(y) = \begin{cases} p + (1-p)f_0(y) & , y = 0 \\ (1-p)f_0(y) & , y \neq 0 \end{cases}$$

- $0 \leq p \leq 1$,
  Y가 원래 가질 수 있는 0을 초과해서 0의 값을 가질 확률
- $f_0(y)$은 영과잉이 발생하기 전의 확률 분포 함수
- 베르누이 분포와 포아송/음이항($f_0(y)$) 분포 혼합모형의 형태

4. 통계적 모델링

모델 평가 방법

- 관심사

상위권 국가의 예측

- 우리의 평가 기준

Y값의 상위 30개에 대한
잔차 절댓값의 평균

$$\frac{\sum_i^{30} |y_i - \hat{y}i|}{30}$$

	y	$\hat{y}$
1	37	27.454
2	34	31.231
3	33	24.683
4	29	31.193
5	28	30.716
⋮		
26	4	2.597
27	4	4.703
28	4	3.491
29	4	0.896
30	4	0.71

4. 통계적 모델링

최종 모델 선택

잔차 절댓값 평균

	Train	Test
선형 회귀	6.701	7.191
포아송 회귀	19.289	18.407
영과잉 포아송	4.631	5.466
영과잉 음이항	6.026	7.512

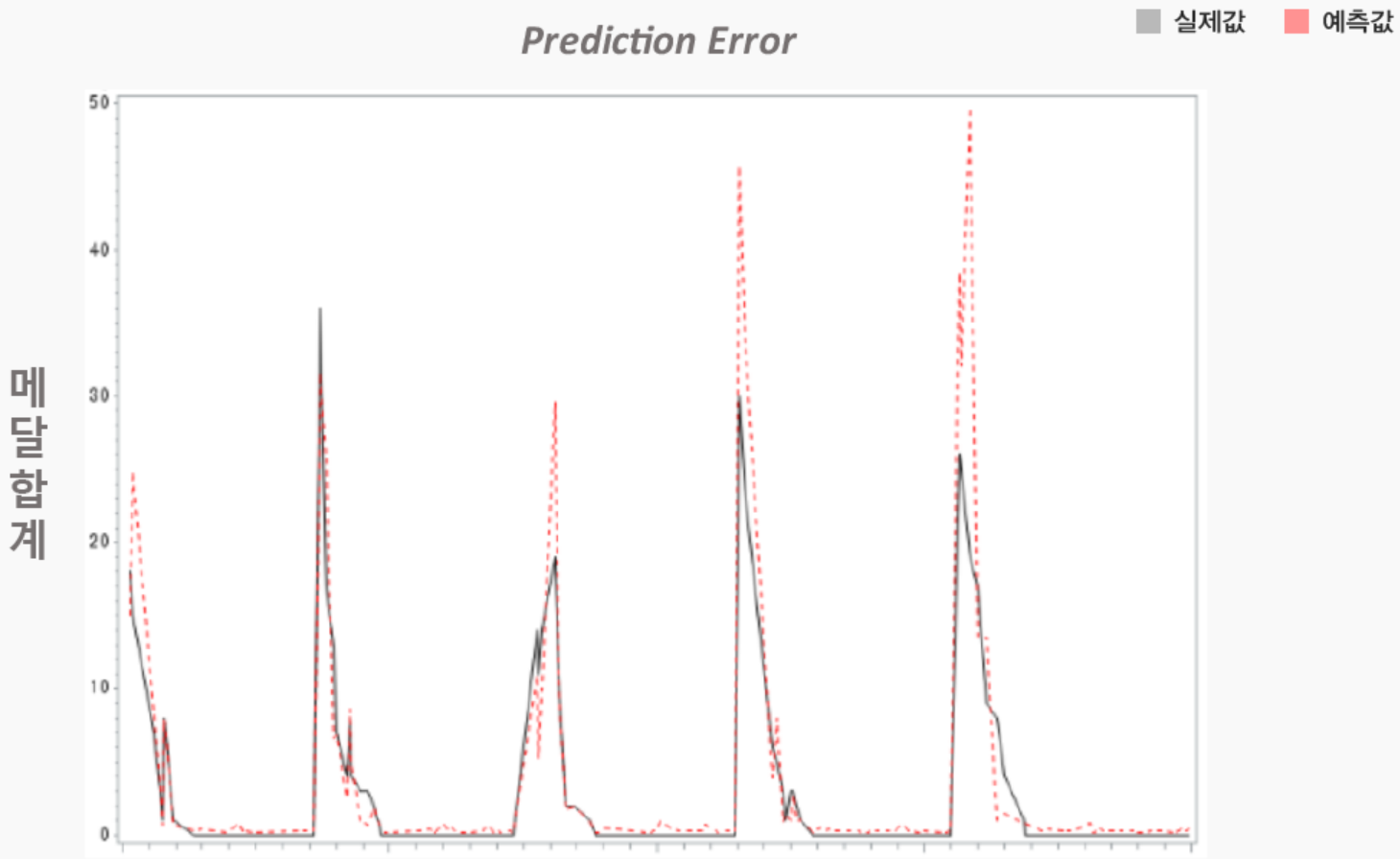
선택 기준

- Test Data의 잔차 절댓값의 평균이 가장 작은 모델
- 영과잉 포아송모델로 최종 선정

4. 통계적 모델링

최종 모델 선택

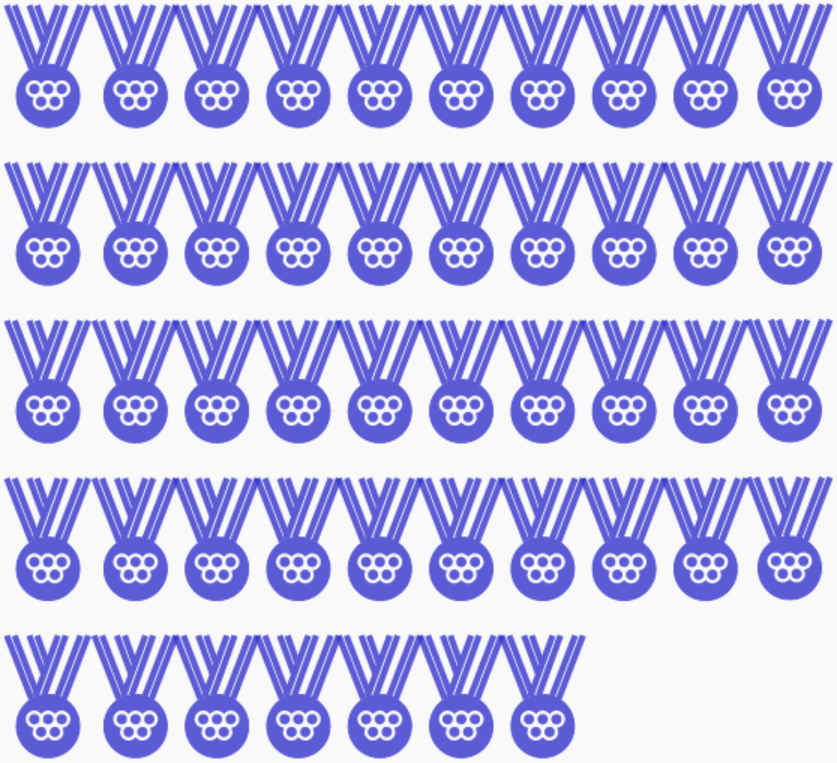
예측값 VS 실제값



팀 소개 데이터 수집 **결론**

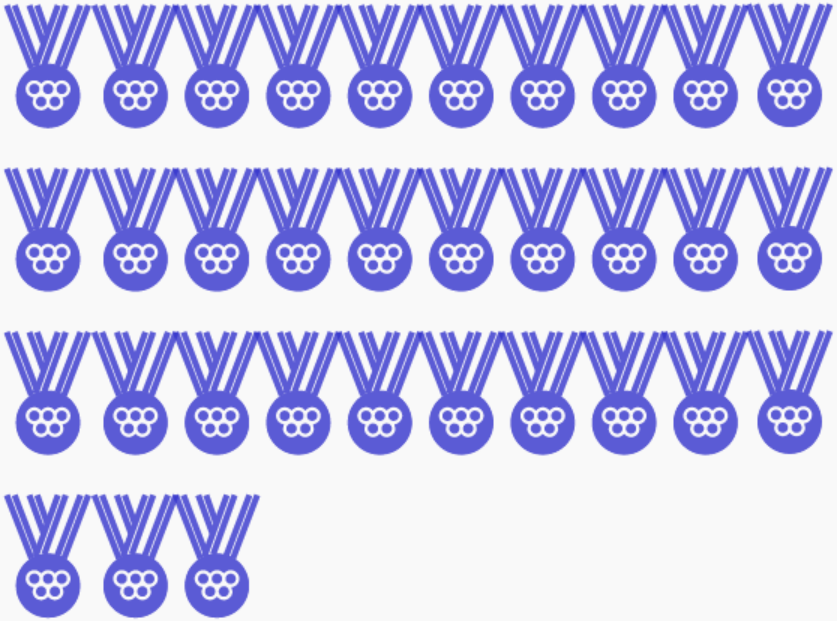
5. 결론 평창 올림픽 메달 예측

독일



5. 결론 평창 올림픽 메달 예측

미국



5. 결론 평창 올림픽 메달 예측

노르웨이

5. 결론 평창 올림픽 메달 예측

캐나다

5. 결론 평창 올림픽 메달 예측

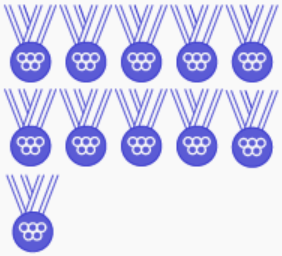
러시아

오스트리아

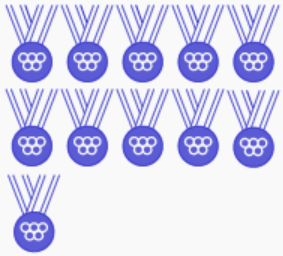
5. 결론 평창 올림픽 메달 예측

스웨덴

중국



프랑스



5. 결론 평창 올림픽 메달 예측

네덜란드

스위스

이탈리아

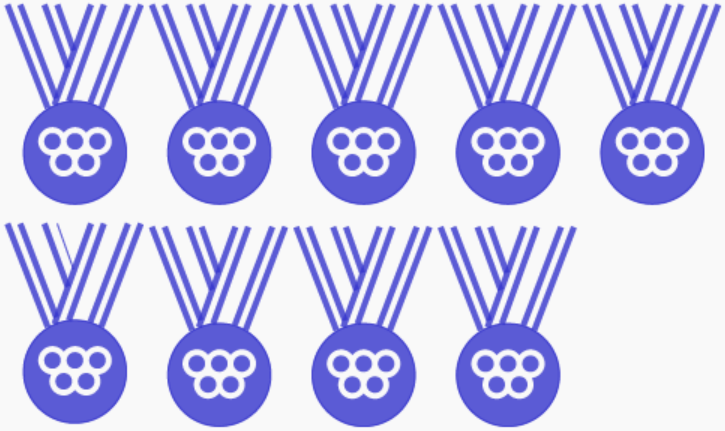
5. 결론 평창 올림픽 메달 예측

핀란드

일본

5. 결론 평창 올림픽 메달 예측

대한민국



주제 선정 통계적 모델링 후기

팀 소개 데이터 수집 결론

후기

“시간의 흐름에 따른 변화 고려하지 못한 것”

“이상치 처리 등 모형을 정교화 시키는 작업 부족”

감사합니다

Q&A