



스토리

경마 데이터 분석

김

성

훈

문

정

승

박

성

은

INDEX

1. 개요

- 경마 소개
- 주제 선정 이유

2. 데이터 수집 및 탐색

- 변수 설명
- 데이터 시각화

3. 데이터 분석

- 적용 모델
- 결과

4. 결론

개요

경마 소개 및 주제 선정 이유

경마 소개



2두 이상의 말이 경주하는 경기에
고객이 돈을 걸어 즐기는 **관람레저스포츠** 중 하나

주제 선정 이유

원래 경마는 사행성 도박이라는 인식이 강해
대중성이 부족했음



렛츠런파크에서 좀비런, 벚꽃축제, 대학생을 대상으로 한 공모전 개최 등으로
다양한 고객이 유입됨. 대중화 성공

최근 인식 변화로
가족, 연인 단위로 찾는 경마장

주제 선정 이유

경마장에 대한 흥미로
데이터 분석 연구를 찾아보았으나,
논문 사례가 많지 않음

- 경마 데이터에 대한 연구나 논문은 가장 최근이 2015년
- 경주마 순위 예측에 대한 논문은 총 3개 뿐



야구, 축구, 농구 등의 스포츠산업에서는
빅데이터 기반 통계 및 예측이 활발히 이루어짐

ex) 야구의 장타율 예측

주제 선정 이유

방식	설명
단승식	경주마 중 1등으로 들어온 말을 적중시키는 방식
연승식	경주마 수에 따라 1,2등 혹은 1,2,3등으로 들어온 말을 적중시키는 방식
복승식	선후착에 관계 없이 1,2등 모두를 적중시키는 방식
쌍승식	선후착 순서를 고려해 1,2등 모두를 적중시키는 방식
복연승식	선후착에 관계 없이 1,2,3등으로 들어온 말 중 두 마리를 적중시키는 방식
삼복승식	선후착에 관계 없이 1,2,3등으로 들어온 말을 모두 적중시키는 방식

선후착에 관계 없이 1, 2, 3두 마를
머신러닝을 이용해 예측하고자 한다.

주제 선정 이유

방식	설명
단승식	경주마 중 1등으로 들어온 말을 적중시키는 방식
연승식	경주마 수에 따라 1,2등 혹은 1,2,3등으로 들어온 말을 적중시키는 방식
복승식	선후착에 관계 없이 1,2등 모두를 적중시키는 방식
쌍승식	“경주마 순위에는 어떤 변수가 영향을 미치는지 분석을 해보자.”
복연승식	선후착에 관계 없이 1,2,3등으로 들어온 말 중 두 마리를 적중시키는 방식
삼복승식	선후착에 관계 없이 1,2,3등으로 들어온 말을 모두 적중시키는 방식

선후착에 관계 없이 1, 2, 3두 마를
머신러닝을 이용해 예측하고자 한다.

데이터 수집 및 탐색

변수 설명 및 데이터 시각화

2. 데이터 수집 및 탐색

데이터 수집

Let's Run 한국마사회 경마정보

서울경마 제주경마 부산경남경마 금주의경마 경마가이드 경마고객의소리 경마방송

공지사향 · 출전정보 · 경마속보 · **경주성적** · 심판정보 · 출발정보 · 경주마정보 · 기수정보 · 조교사정보 · 마주정보 · 자료실 · 경마시행정보 · KHC

서울경마

경주성적

경주성적표 > 조건별 경주성적표 > 역대 대상경주 우승마 > 대상경주 갤러리 > 역대 연도 대표마 > 거리별 최고기록 > 등급/거리별기록 > 조건별 최고 평균기록 > 승석별 최고 평균배당 > PDF파일 > 경주결과 기록지 > 역대 시리즈 경주성적 >

경주성적표

경주성적표 리스트 일자별 요약성적표 **경주별 상세 성적표**

경주별 상세 성적표 단 · 연 · 복 경승식 복면승식 삼복승식 삼평승식

경주영상 고화질 모바일 경주전 영상 예시장 경주로입장 < 이전경주 다음경주 > 인쇄하기 > 역선택화 >

2019년 05월 05일 (일) 제1경주 서울 제34일 맞을 건조 (5%) 10:45

국6등급 1000M 별정A 일반 R0~0 연령 오픈

순위상금

13,680,000원

5,040,000원

3,120,000원

1,200,000원

960,000원

마명앞에 ★표시는 골번신청됩니다.

중량앞에 *표시는 52kg미만의 말이 52kg으로 중량신청한 것을 의미합니다.

마주명앞에 ◆표시는 마주복색 등록마주를 의미합니다.

장구현황의 +는 신규장구, -는 해지장구를 의미합니다. (2016년 2월 성적표자료부터 표기합니다.)

순위	마번	마명	산지	성별	연령	중량	레이팅	기수명	조교사명	마주명	도착마	마체중	단승	연승	장구현황
1	2	인디펜던스	한	수	3세	56		김홍근	송문길	(주)나스카		485(11)	1.8	1.1	
2	11	인양레오파트	한	암	4세	53		(-)김효정	구자홍	(주)에이애프피	3	417(-8)	15.9	2.7	망사
3	9	스틸런	한	암	4세	54		이동하	박윤규	박덕희	1	433(10)	99.2	10.0	망사돈
4	4	코리아대장부	한	거	3세	56		황순도	배휴준	곽승홍	±	510(-1)	63.4	10.2	계란철근, 허곤, 눈가면
5	8	서울매직	한	수	3세	53		(-)문성현	김동철	◆이시환	코	455(0)	5.0	1.4	
6	7	신의축복	한	암	3세	54		문로	심승태	김교환	3	458(2)	20.2	3.6	망사돈
7	3	지오바나	한	암	3세	54		함완식	박재우	오종혁	1	467(-9)	4.2	1.6	망사돈
8	12	서던맨	한	거	3세	56		마누엘	홍대유	◆이경희	±	461(-4)	140.8	19.7	
9	1	선사인특급	한	암	3세	54		유승환	이신영	박의주	1	484(2)	31.3	4.7	계란철근, 망사
10	6	함프윈	한	암	3세	54		다비드	이희영	◆서순배	±	444(-1)	119.6	24.2	망사

Let's Run Studbook 한국마사회 말혈통정보

마명, 마번

검색

말정보 말 등록 씨수말 정보 교배정보 경매 혈통경마 세계의 경마 유전자정보

개별말정보조회 등록마내역 수출마내역 수입마내역 경주마 철고내역 육성훈련장사 출생및기타정보

HORSE INFORMATION 세대를 거슬러 선조들의 자취를 되찾아보라

말정보 > 개별말정보조회 > 마적사항

마적사항

마적사항

말등록부

부개보

모개보

소유변동

용도변동

소재변동

친로내역

방역내역

경매내역

육성훈련

훈련장사

훈련내역

경주성적

연방발상적

출전정보

군번동

등록신청내역

개체식별

프로필

육종가

2 마적사항

생산지

혈통서

혈통등록일

부마명

모마명

경주성적

외국경주

(주)북원목장(북원목장)

2017-01-07

윤로실버

외전한디시

5전(1.0/1.1/1.0)

말명

일마레 (IL MARE)

2018-07-13

수

더러브렛

한국

경주용

소유자

강도노 시계요

조교사

선범석(12조)

2018-07-20 ~

생산국

미국

생산국

미국

수득상금

19,800,000원

수득상금

마번

말이레 (IL MARE)

2016-07-13

출생일

2016-05-20 (3세)

말색

갈색

종종

국제마인

KOR16139565

생산국

수입국

원마명

소제지

서울경마공회

패스포트

2017-03-16

번식등록일

2001

생년도

2001

국(등급)

국5

우승거리

사진 출처 : 한국 마사회, 한국 마사회 말 혈통정보

크롤링을 통해
한국 마사회 사이트, 한국마사회 말혈통정보 사이트에서
약 2년치(2017.01~2019.04) 데이터를 수집, 약 28000개의 자료

2. 데이터 수집 및 탐색

데이터 수집

순위	마번	마명	산지	성별	연령	중량	기수명	조교사명	마주명	마체중	단승	연승	장구현황	구간별순위	S-1F	1코너	2코너	3코너	4코너
1	8	일마레	한	수	3	55.5	먼로	서범석	카도노 시가	472	18.6	3.8	계란형큰,	2- - - 2	00:13.5			00:30.2	00:48.3
2	6	천하여제	한	암	3	53	문세영	지용철	금악목장	493	3.1	1.4		1- - - 1	00:13.4			00:30.2	00:48.2
3	4	백두거포	한	거	3	55	마누엘	구자흥	김봉겸	472	10.8	3	계란형큰,	10- - -	00:14.0			00:30.8	00:49.0
4	12	쏘아라흑조	한	수	3	55	이준철	김대근	현대봉	545	66.8	18		9- - - 1	00:13.9			00:30.9	00:49.5
5	11	퍼스트선더	한	수	4	55	김정준	심승태	장철환	491	45.3	6.1	망사, 반가	8- - - 9	00:13.7			00:30.9	00:49.5
6	5	대호천하	한	거	4	55	정평수	김동철	고재완	463	34.1	7.1	계란형큰,	6- - - 8	00:13.7			00:30.8	00:49.0
7	2	달리는웅지	한	수	4	56	유승완	박희철	경규환	477	4.9	1.5	계란형큰,	4- - - 4	00:13.6			00:30.5	00:48.8
8	1	캐럿시커	한	수	4	55.5	함완식	김동균	공이공팔	454	4.2	1.6	계란형큰+	3- - - 3	00:13.6			00:30.5	00:48.8
9	7	선샤인마린	한	암	3	54	다나카	이관호	이경희	482	20.1	4.1	계란형큰,	12- - -	00:14.3			00:31.9	00:50.6
10	10	아임해피	한	암	4	54	빅투아르	박재우	김정철	455	5.3	2.6	망사눈	7- - - 6	00:13.7			00:30.7	00:49.0
11	3	블랙스피릿	한	암	4	54	다비드	김학수	이혜란	434	42.6	9.5	망사눈+,Ti	5- - - 5	00:13.6			00:30.5	00:48.8
12	9	매직돌풍	한	암	4	53.5	안효리	유재길	오왕근	471	86.1	21.2	반가지큰,	11- - -	00:14.0			00:31.5	00:50.4
1	8	럭키봉성	한	암	3	53.5	김동수	박재우	최기영	502	2.6	1.3	망사, 반가	1- - - 1	00:13.1			00:13.1	00:29.4
2	6	미스겔라	한	암	3	52	문세영	배휴준	김창호	447	3.2	1.4	계란형큰+	4- - - 4	00:13.5			00:13.5	00:30.4
3	2	엘초이스	한	암	4	54.5	최범현	강환민	이종욱	492	7.6	2.4	반가지큰,	2- - - 2	00:13.3			00:13.3	00:29.7
4	10	레이징엔젤	한	암	4	53.5	임기원	박종곤	한종희	484	12.4	2.3	반가지큰,	3- - - 3	00:13.3			00:13.3	00:29.7
5	3	금스타	한	암	3	52.5	김정준	박병일	배태곤	488	6.1	2	망사눈	6- - - 6	00:13.9			00:13.9	00:30.7
6	4	흥깨비	한	암	3	53	유승완	최상식	고광숙	470	11.1	2.1	양털코	10- - -	00:14.2			00:14.2	00:31.1
7	1	맹산히어로	로미	수	3	55.5	이동하	지용철	조용학	488	24.6	5.1	Triabit	9- - - 9	00:14.2			00:14.2	00:31.1
8	7	기븐러너	한	수	3	53	조상범	안병기	김종철	470	38.2	5.1	망사눈	5- - - 5	00:13.6			00:13.6	00:30.4
9	9	아이엠위너	한	암	4	51	방춘식	박흥진	오왕근	458	135.4	27.5	망사눈	8- - - 8	00:14.1			00:14.1	00:31.8
10	5	헌터호크	한	거	5	51.5	이동진	박천서	윤훈현	500	79.2	15.1	혀끈	7- - - 7	00:14.1			00:14.1	00:31.6

변수 설명

요인	설명
순위	경주마가 들어온 순서
마번	경주시 말에게 부여되는 번호
마명	경주마의 이름
산지	경주마의 산지
성별	경주마의 성별로 암컷, 수컷, 거세로 구분
연령	경주마의 나이
중량	기수의 체중
기수명	경주마를 타는 기수의 이름
조교사명	경주마를 관리하는 사람의 이름
마주명	경주마의 주인
마체중	경주마의 체중
단승	선택한 말이 1착 할 때의 배당
연승	선택한 말이 1착 또는 2착 할 때의 배당
장구현황	경주마가 경주에서 짊어지게 되는 장구류
구간별순위	경주마가 구간별로 들어온 순위

2. 데이터 수집 및 탐색

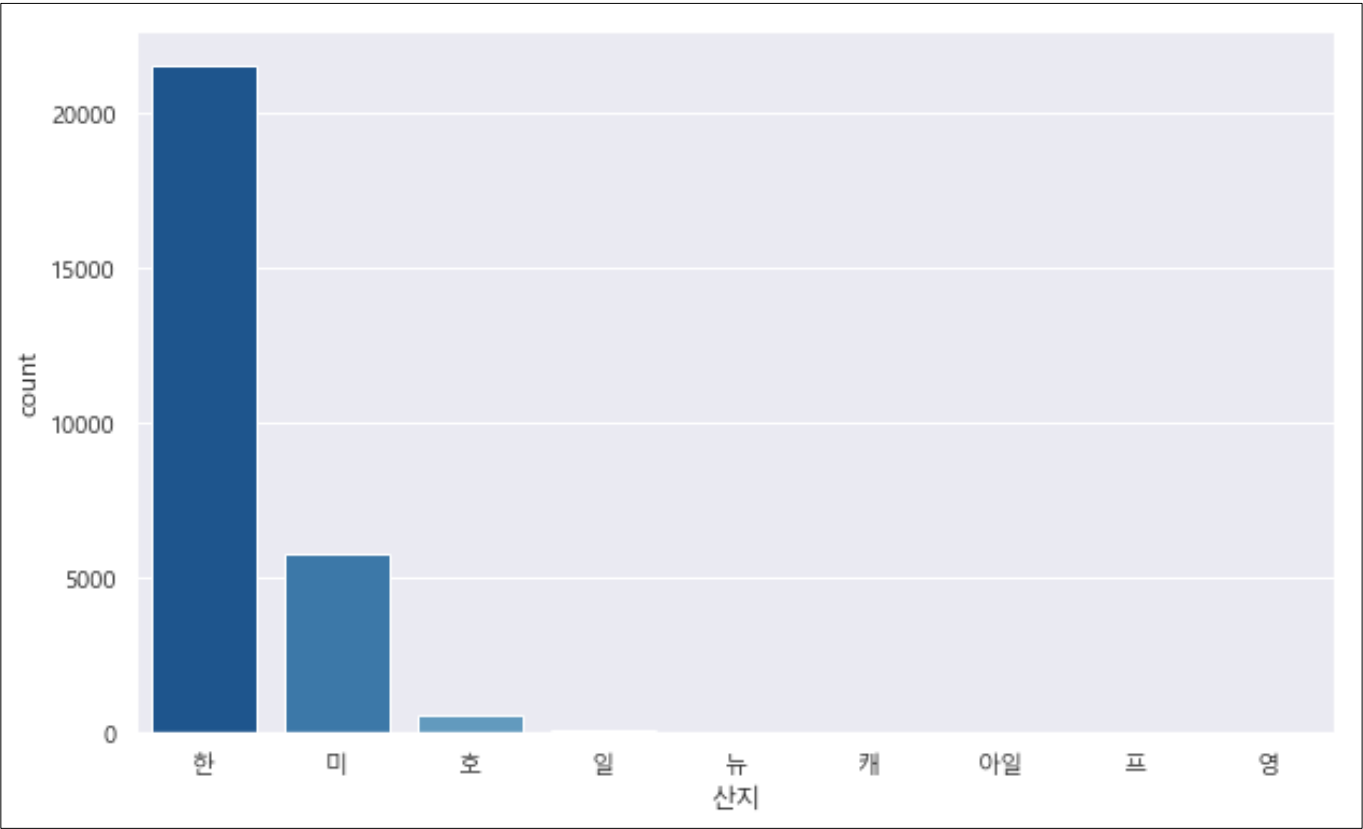
변수 설명

요인	설명
S-1F	출발선에서 200m까지 기록
1코너	첫번째 코너까지 걸리는 시간
2코너	두번째 코너까지 걸리는 시간
3코너	세번째 코너까지 걸리는 시간
4코너	네번째 코너까지 걸리는 시간
G-3F	결승선에서 600m 직전까지 기록
G-1F	결승선에서 200m 직전까지 기록
경주기록	한 경기의 기록
날짜	경마 경주가 시행된 날짜
날씨	경마 경주가 시행된 시점의 날씨
주로상태	경마 경주가 시행된 시점의 주로 상태
주로습도	경마 경주가 시행된 시점의 주로 습도
등급	경주마가 속해 있는 등급
거리	경마경주시 경주마가 뛴 거리로 단위는 M

“ 총 29개의 변수 ”

데이터 탐색

산지 경주마의 산지

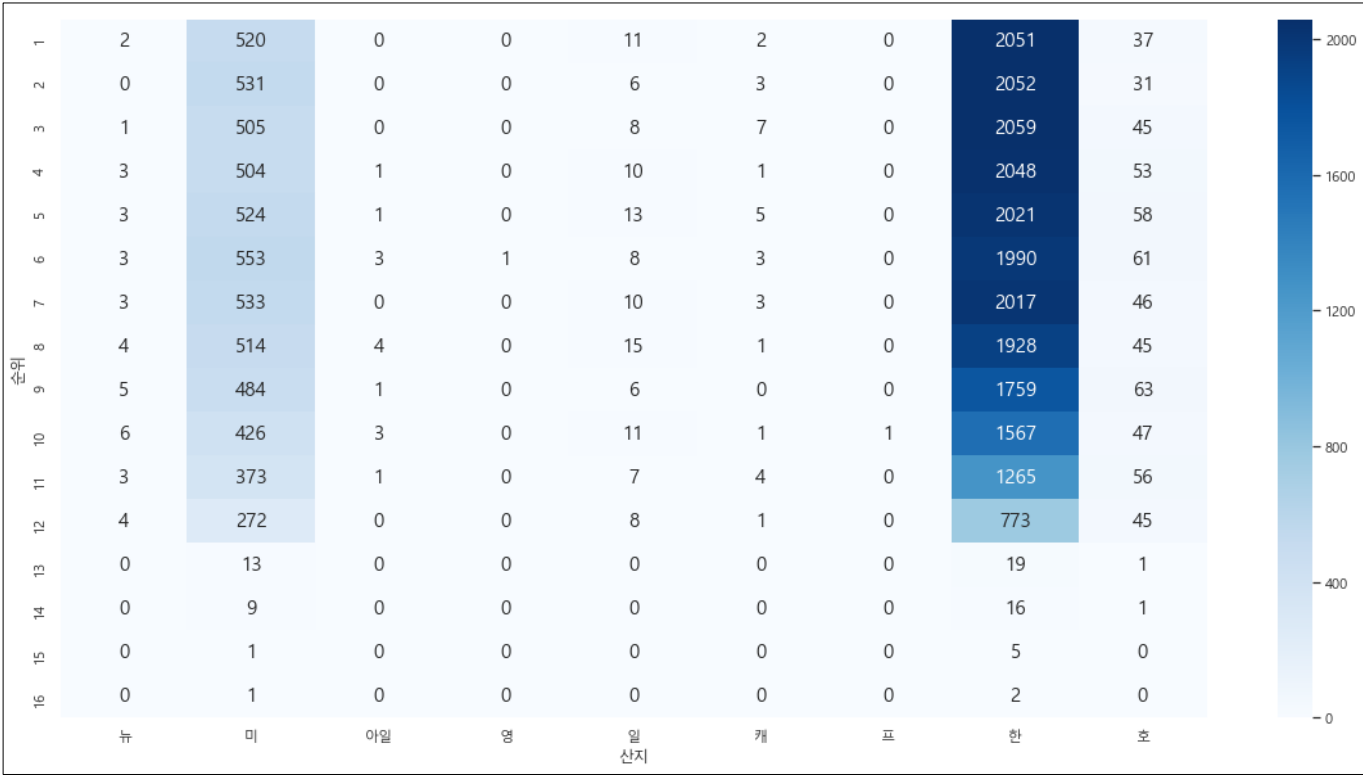


산지	빈도
한국	21572
미국	5763
호주	589
일본	113
뉴질랜드	37
캐나다	31
아일랜드	14
프랑스	1
영국	1

2. 데이터 수집 및 탐색

데이터 탐색

산지 & 순위 연관성



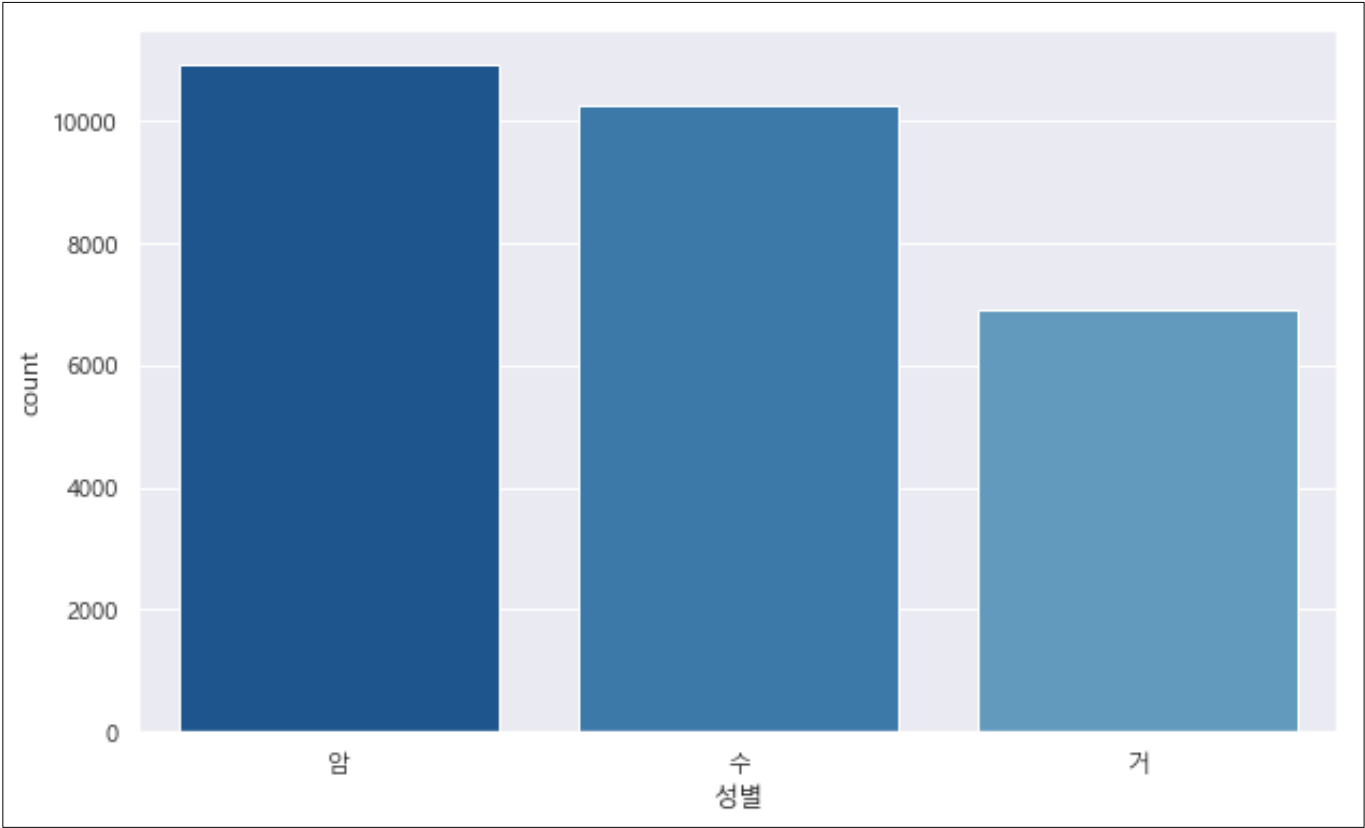
H0: 산지와 순위는 독립이다.
H1: 산지와 순위는 독립이 아니다.

카이제곱 검정			
	값	자유도	점근유의확률 (양측검정)
Pearson 카이제곱	182.257 ^a	120	.000
우도비	168.803	120	.002
유효 케이스 수	28121		

유의 확률 < 0.05이므로, 귀무가설 기각
즉, 산지와 순위는 독립이 아니다.

데이터 탐색

성별 경주마의 성별로 암컷, 수컷, 거세로 구분

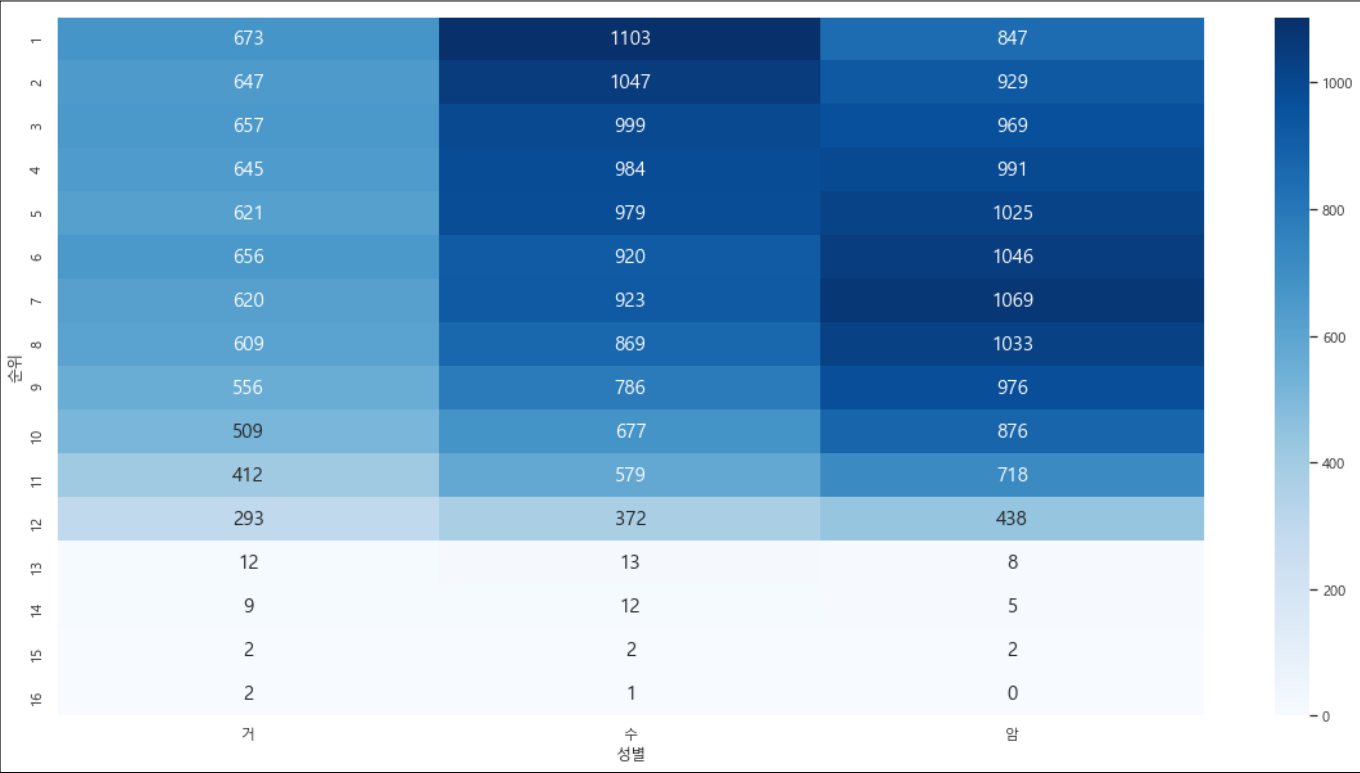


성별	빈도
암컷	10932
수컷	10266
거세	6923

2. 데이터 수집 및 탐색

데이터 탐색

성별 & 순위 상관성



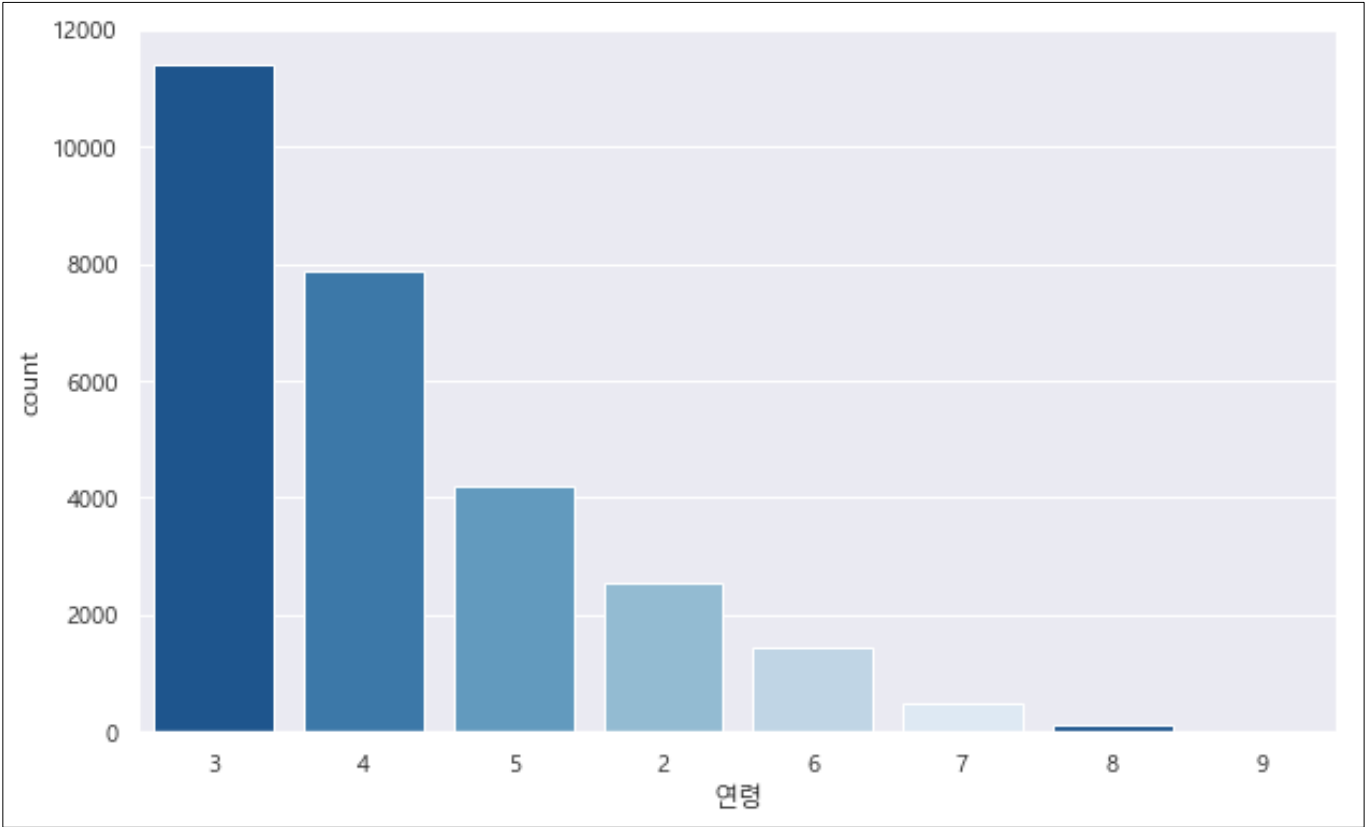
H0: 성별과 순위는 독립이다
H1: 성별과 순위는 독립이 아니다.

카이제곱 검정			
	값	자유도	점근유의확률 (양측검정)
Pearson 카이제곱	138.281 ^a	30	.000
우도비	139.950	30	.000
유효 케이스 수	28121		

유의 확률 < 0.05이므로, 귀무가설 기각
즉, 성별과 순위는 독립이 아니다.

데이터 탐색

연령 경주마의 나이

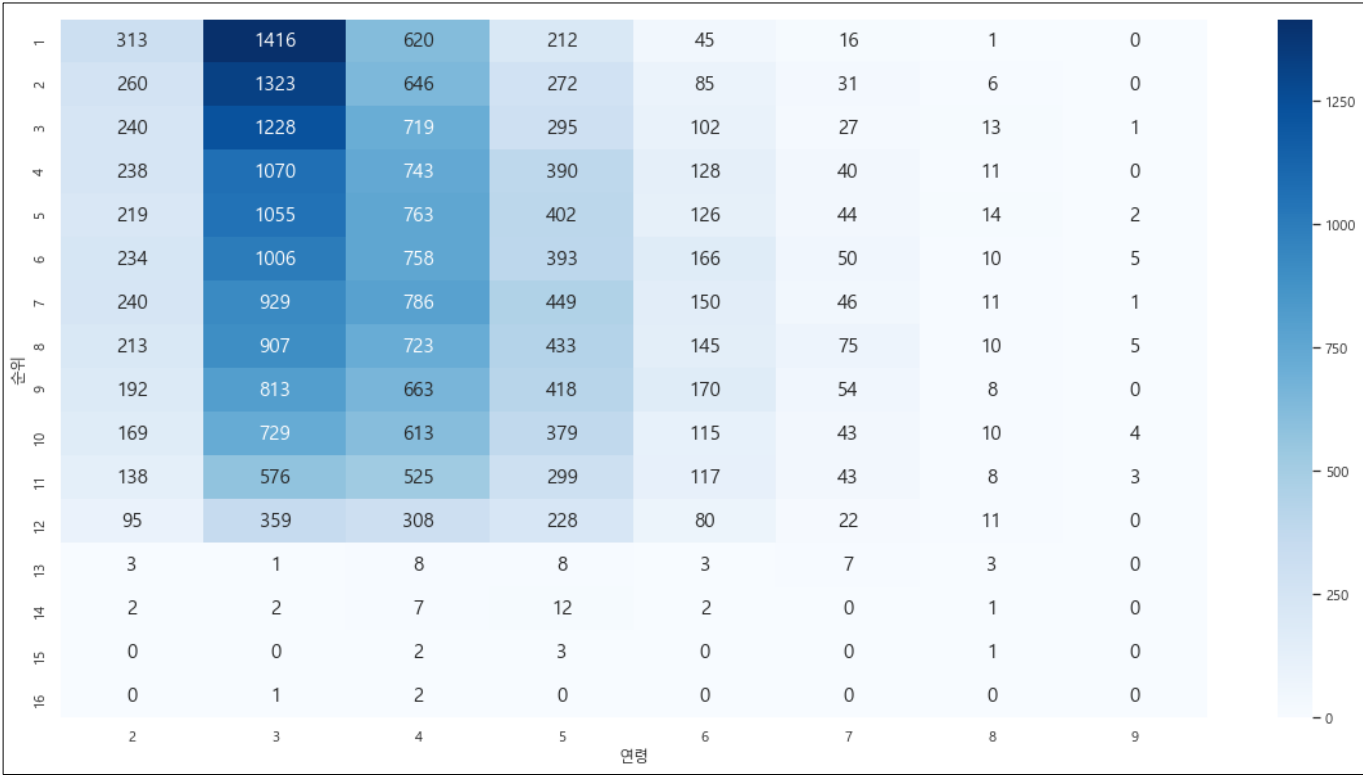


연령	빈도
3세	11415
4세	7886
5세	4193
2세	2556
6세	1434
7세	498
8세	118
9세	21

2. 데이터 수집 및 탐색

데이터 탐색

연령 & 순위 상관성



H0: 연령과 순위는 독립이다
H1: 연령과 순위는 독립이 아니다.

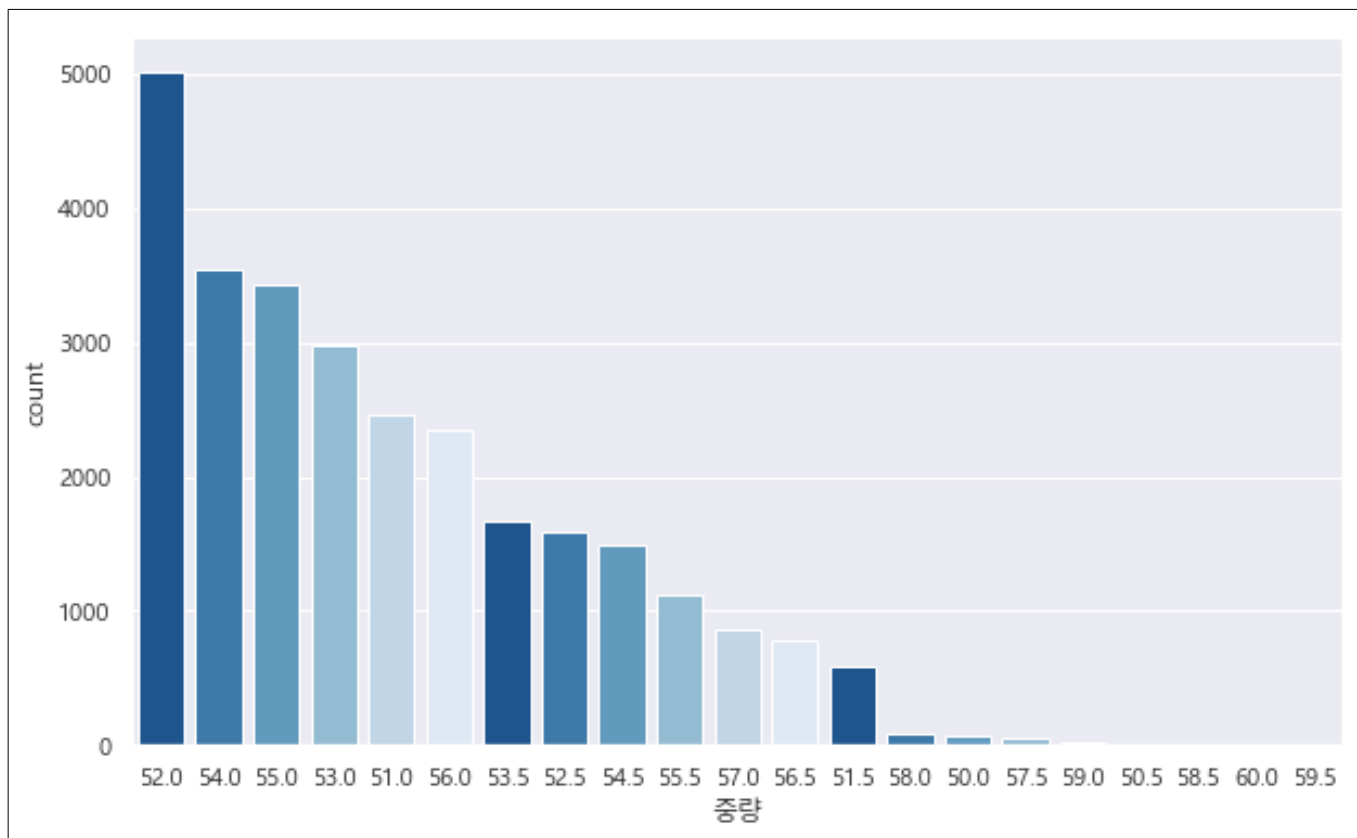
카이제곱 검정			
	값	자유도	점근유의확률 (양측검정)
Pearson 카이제곱	1084.902 ^a	105	.000
우도비	995.790	105	.000
선형 대 선형결합	597.313	1	.000
유효 케이스 수	28121		

유의 확률 < 0.05이므로, 귀무가설 기각
즉, **연령과 순위는 독립이 아니다.**

2. 데이터 수집 및 탐색

데이터 탐색

중량 기수의 체중

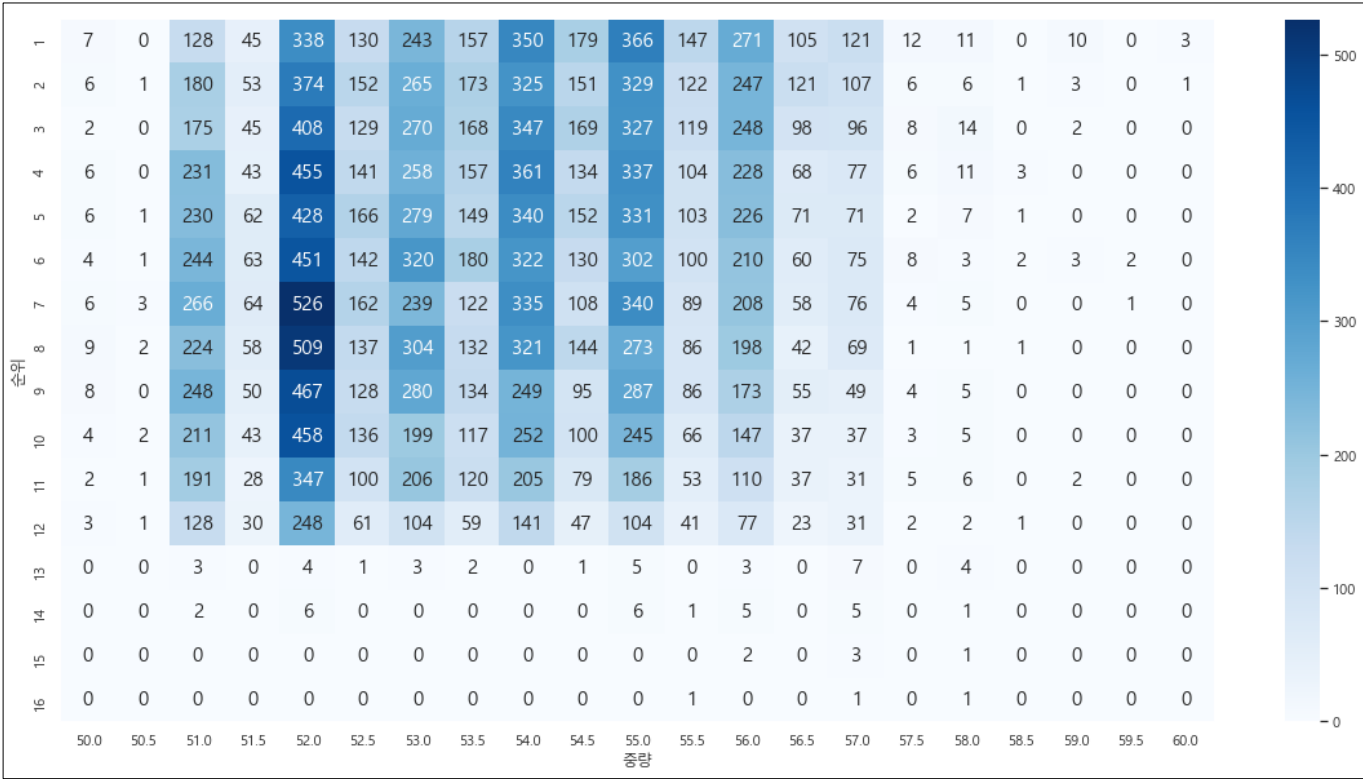


[52kg의 기수가 가장 많음을 확인]

2. 데이터 수집 및 탐색

데이터 탐색

중량 & 순위 상관성



H0: 중량과 순위는 독립이다
H1: 중량과 순위는 독립이 아니다.

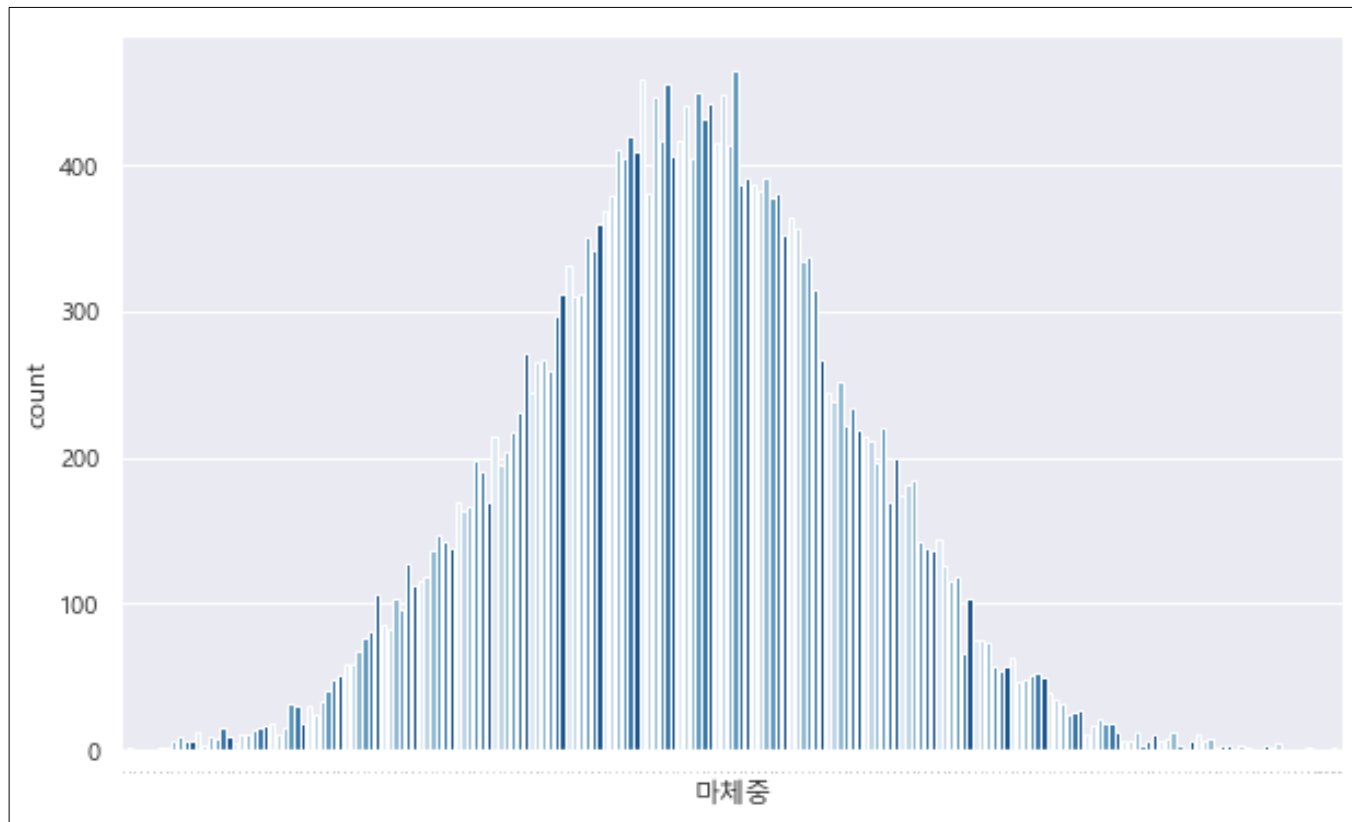
카이제곱 검정			
	값	자유도	점근유의확률 (양측검정)
Pearson 카이제곱	1263.476 ^a	300	.000
우도비	910.250	300	.000
선형 대 선형결합	383.313	1	.000
유효 케이스 수	28121		

유의 확률 < 0.05이므로, 귀무가설 기각
즉, 중량과 순위는 독립이 아니다.

2. 데이터 수집 및 탐색

데이터 탐색

마체중 경주마의 체중

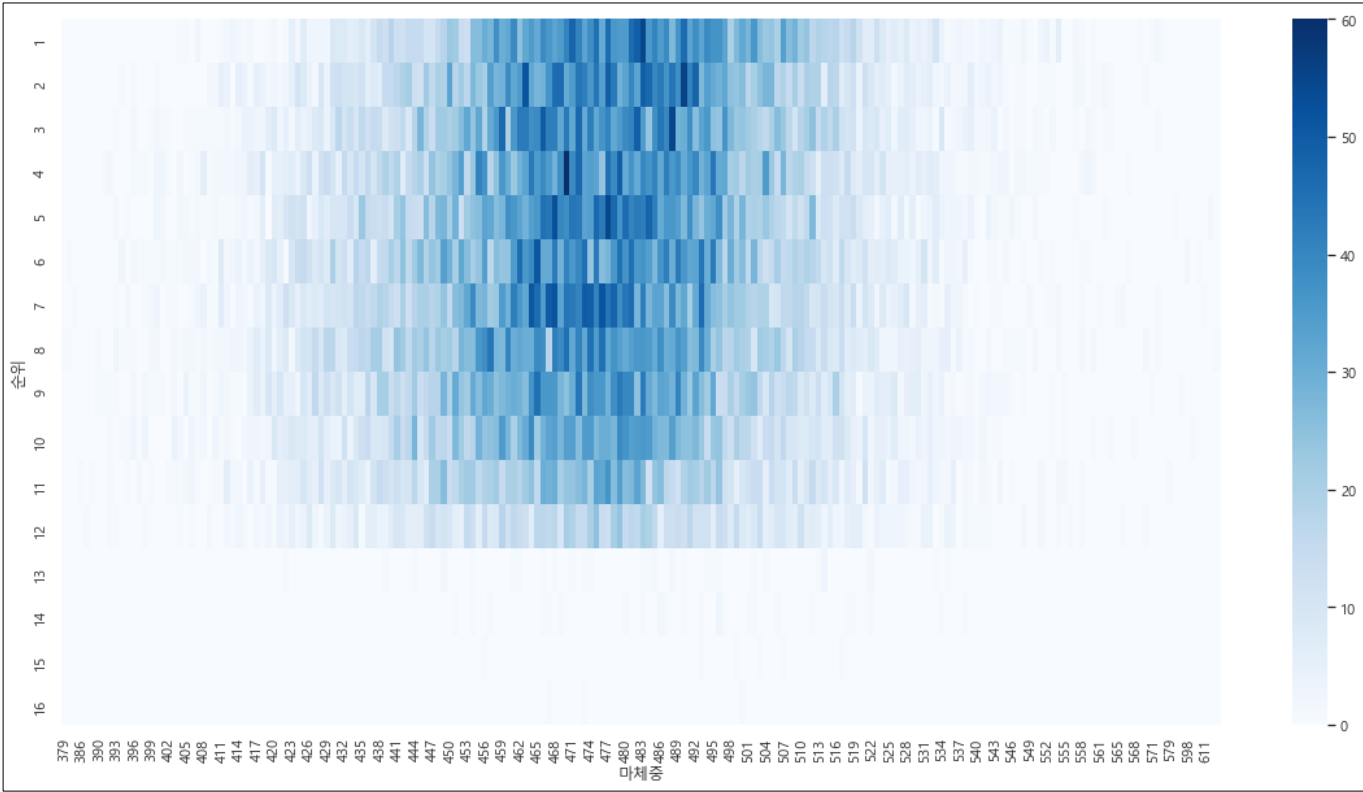


[480kg을 중심으로 퍼져 있음을 확인]

2. 데이터 수집 및 탐색

데이터 탐색

마체중 & 순위 상관성



H0: 마체중과 순위는 독립이다
H1: 마체중과 순위는 독립이 아니다.

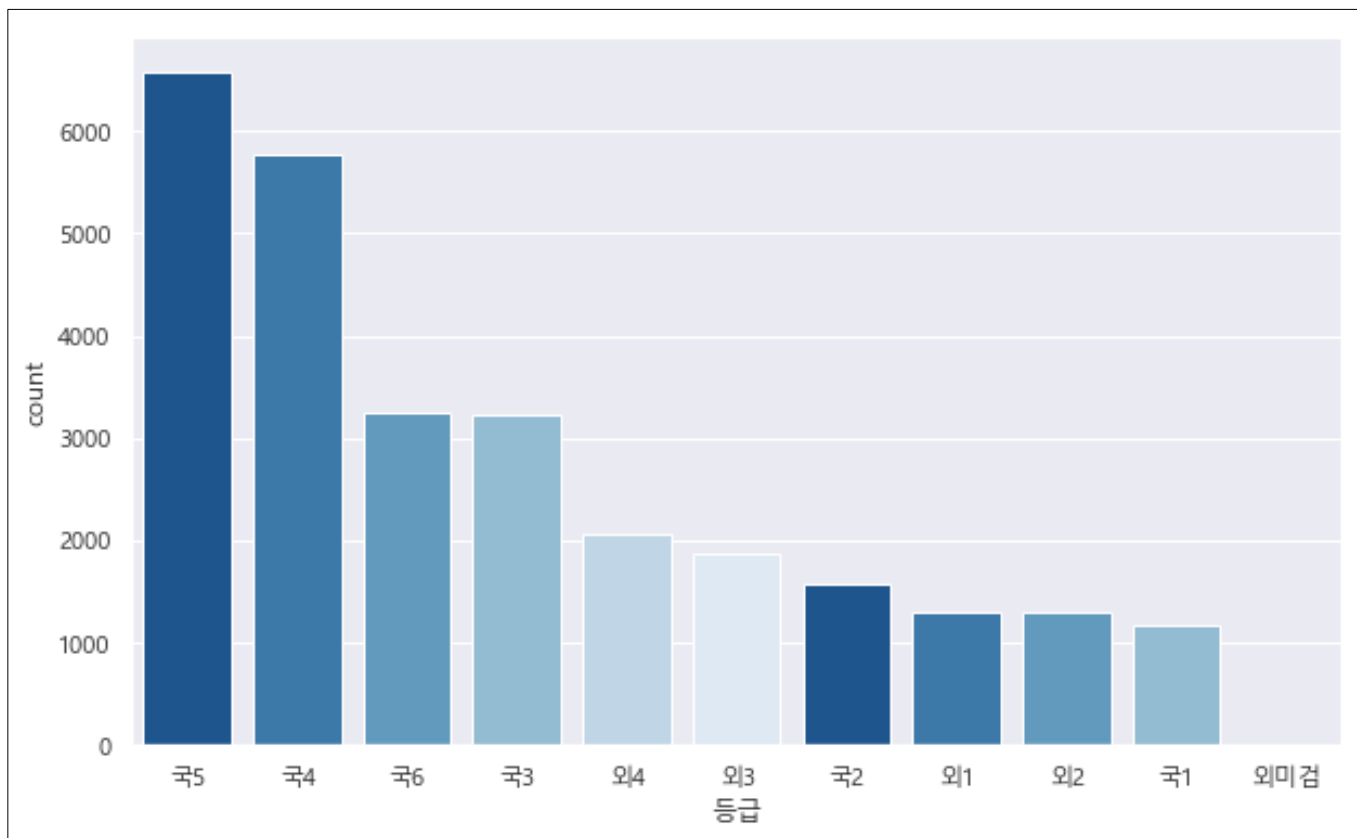
카이제곱 검정			
	값	자유도	점근유의확률 (양측검정)
Pearson 카이제곱	3226.647 ^a	2955	.000
우도비	2788.773	2955	.986
선형 대 선형결합	135.955	1	.000
유효 케이스 수	28121		

유의 확률 < 0.05이므로, 귀무가설 기각
즉, 마체중과 순위는 독립이 아니다.

2. 데이터 수집 및 탐색

데이터 탐색

등급 경주마가 속해 있는 등급

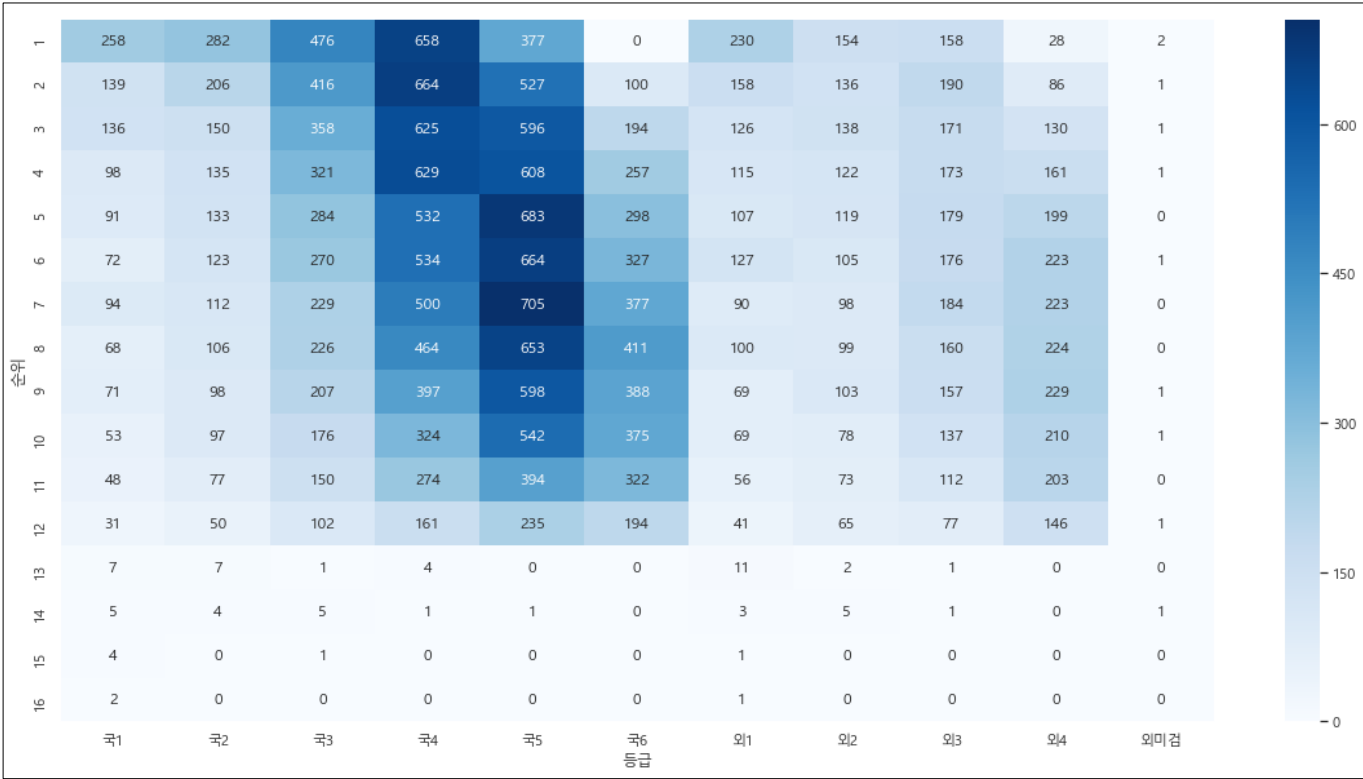


등급	빈도
국내 5등급	6583
국내 4등급	5767
국내 6등급	3243
국내 3등급	3222
외국 4등급	2062
외국 3등급	1876
국내 2등급	1580
외국 1등급	1304
외국 2등급	1297
국내 1등급	1177
외국산 미검정	10

2. 데이터 수집 및 탐색

데이터 탐색

등급 & 순위 상관성



H0: 등급과 순위는 독립이다
H1: 등급과 순위는 독립이 아니다.

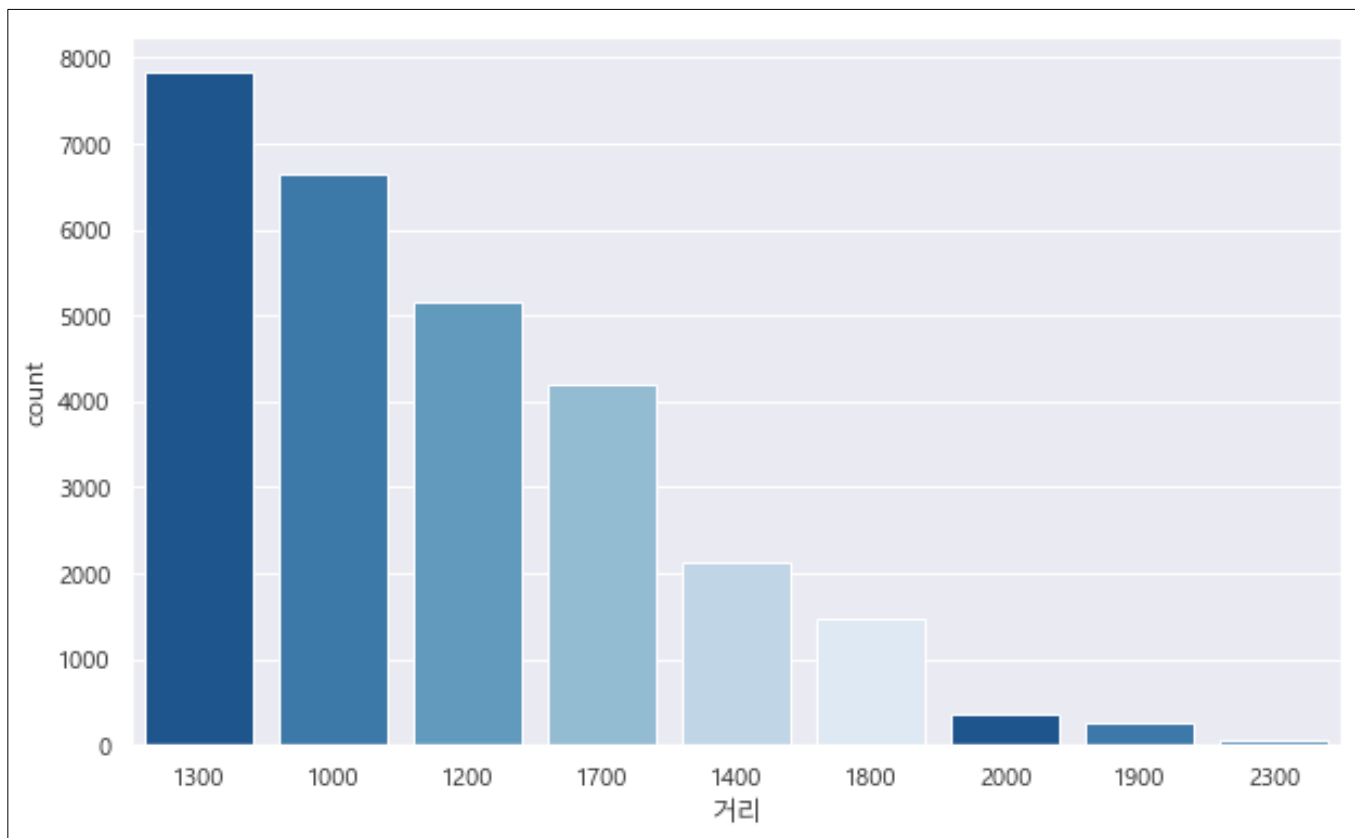
카이제곱 검정			
	값	자유도	점근유의확률 (양측검정)
Pearson 카이제곱	2767.894 ^a	150	.000
우도비	2869.587	150	.000
유효 케이스 수	28121		

유의 확률 < 0.05이므로, 귀무가설 기각
즉, 등급과 순위는 독립이 아니다.

2. 데이터 수집 및 탐색

데이터 탐색

거리 경마경주시 경주마가 뛴 거리로 단위는 M



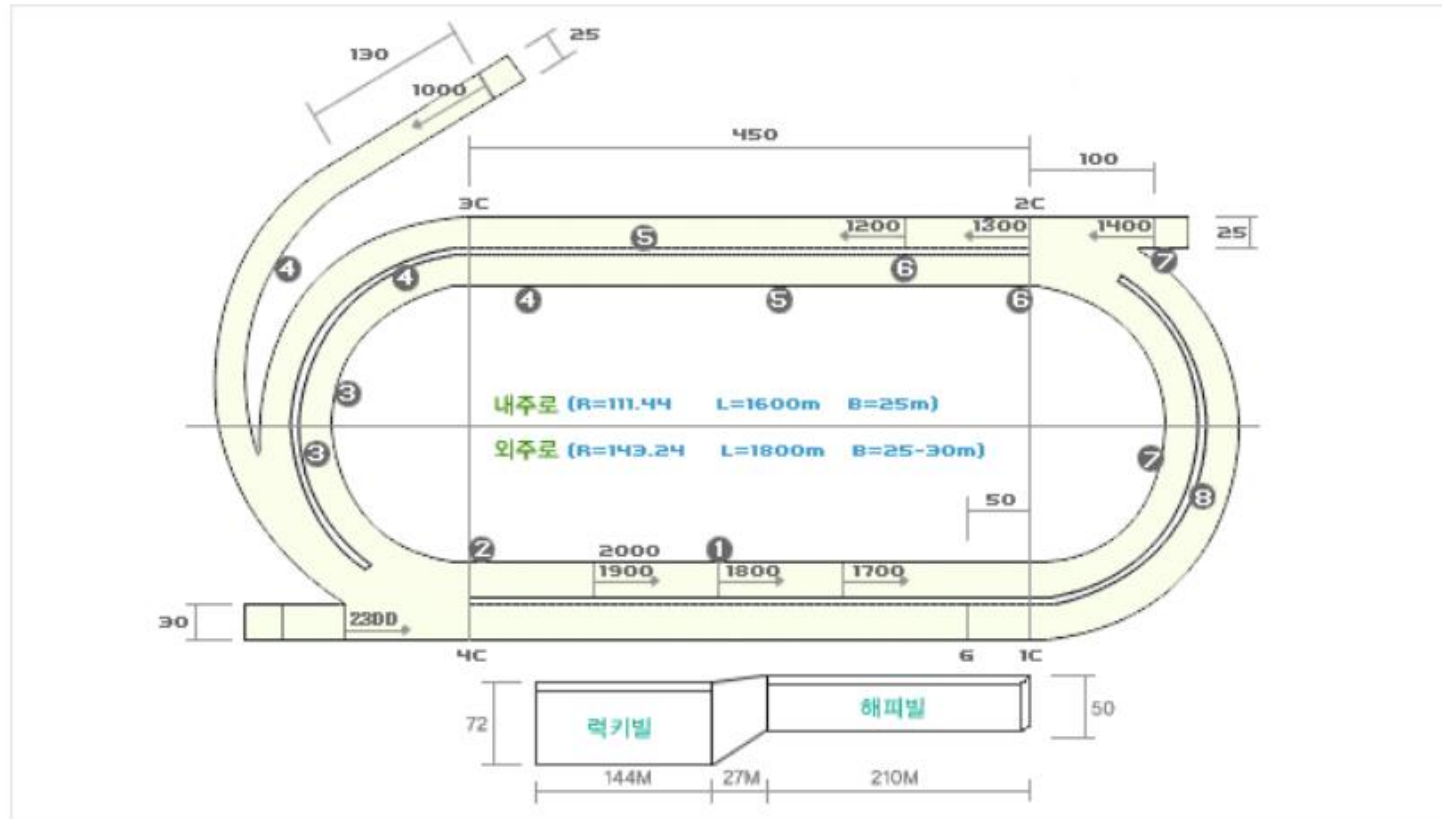
거리	빈도
1300M	7841
1000M	6634
1200M	5149
1700M	4201
1400M	2127
1800M	1481
2000M	368
1900M	263
2300M	57

2. 데이터 수집 및 탐색

데이터 탐색

마번 경주시 말에게 부여되는 번호

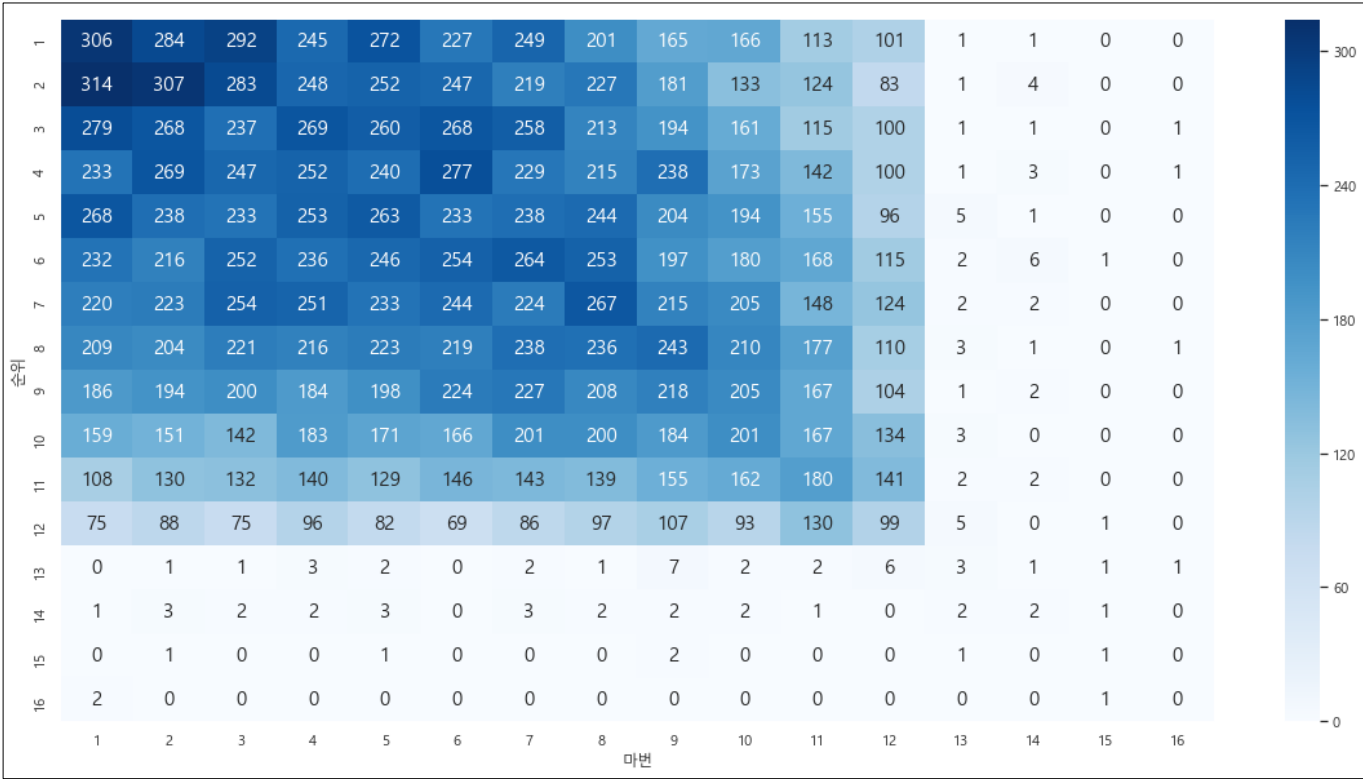
■ 경주로 평면도



2. 데이터 수집 및 탐색

데이터 탐색

마번 & 순위 상관성



H0: 마번과 순위는 독립이다
H1: 마번과 순위는 독립이 아니다.

카이제곱 검정			
	값	자유도	점근유의확률 (양측검정)
Pearson 카이제곱	4342.54 ^a	225	.000
우도비	875.743	225	.000
선형 대 선형결합	483.383	1	.000
유효 케이스 수	28121		

유의 확률 < 0.05이므로, 귀무가설 기각
즉, 마번과 순위는 독립이 아니다.

순위는

말의 산지, 성별, 연령, 체중, 등급과
기수의 중량, 경기의 거리, 마번에

영향을 받음을 확인

데이터 분석

적용 모델 및 결과

3. 데이터 분석

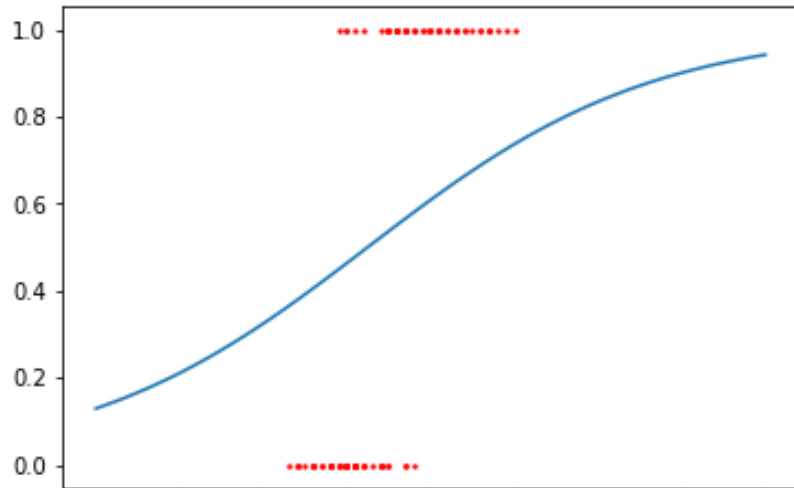
적용 모델

Logistic Regression

정의

- 독립 변수들의 선형 결합을 이용하여 사건의 발생 가능성을 예측하는 모델

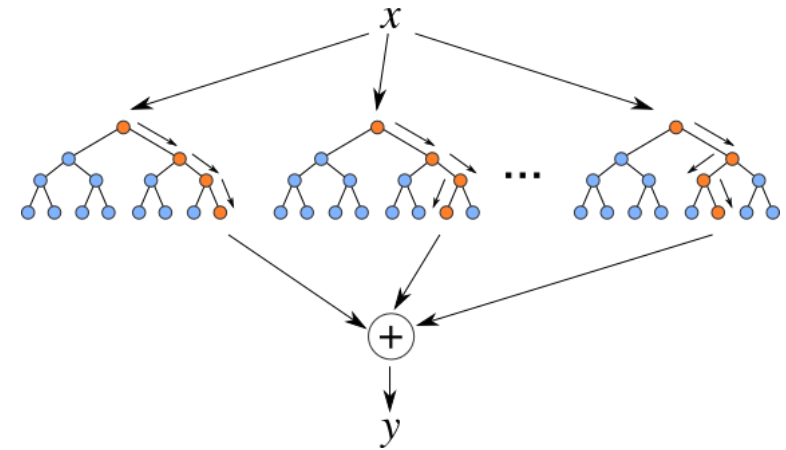
$$\Pr(Y = 1|x) = p(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$



Random Forest

정의

- o/x로 의사 결정을 내려주는 의사 결정 트리를 여러 개 생성하여 투표를 통해 결정을 내려주는 모델



3. 데이터 분석

적용 모델

Logistic Regression 사건의 발생 가능성을 모델링을 통해 판별 후 일정 기준으로 분류

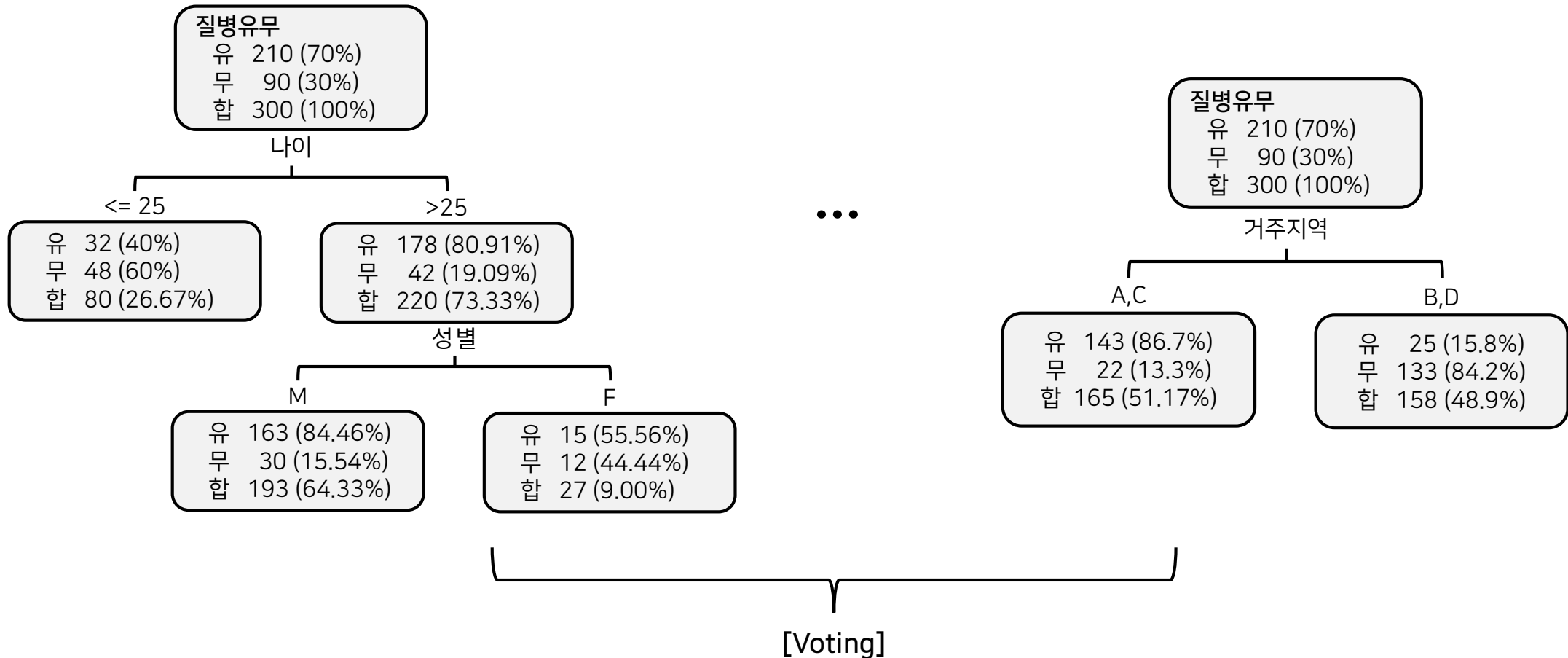
나이	성별	거주지역	질병 유무		예측 질병 확률		예측 질병 유무
X1	x2	X4	y		P(y=1)		$\hat{y}(0.50)$
33	F	A	0	[Discriminant] →	0.35	[Classification] →	0
35	M	D	1		0.93		1
45	F	B	0		0.73		1
27	F	C	1		0.15		0
29	M	A	1		0.56		1

$$\hat{p}(y = 1) = \frac{\exp(a + b_1x_1 + b_2x_2 + \dots + b_px_p)}{1 + \exp(a + b_1x_1 + b_2x_2 + \dots + b_px_p)}$$

3. 데이터 분석

적용 모델

Random Forest 각기 다른 의사결정나무를 여러 개 생성하여 학습시켜 다수결 방식으로 결과 도출



모델 학습 준비

- 데이터 탐색과 기존 연구 조사를 통해 유의미하다고 판단된 17개의 변수들로 학습을 진행

학습 변수
순위, 마번, 연령, 종량, 마체중, 단승, 연승, 주로습도, 거리, 코너, S1F, G3F, G1F, 등급, 성별, 날씨, 기수명

- 결측치를 제외한 26828개의 데이터 사용
- Training: 80% / test: 20% / 30번 반복 실험
- 1마 - 경주마 중 1등으로 들어온 말을 적중
- 2마 - 선후착에 관계 없이 1,2등으로 들어오는 말 적중
- 3마 - 선후착에 관계 없이 1,2,3등으로 들어오는 말 적중

3. 데이터 분석

모델 학습 결과

학습 결과

데이터 불균형 문제를 해결하기 위해 클래스 빈도에 반비례하는 가중치를 두고 학습

Logistic Regression

1마		예측결과	
		False	True
실제 정답	False	4844	37
	True	442	43

Accuracy: 91%
Recall(1): 9%
Precision(1): 54%
F1-score(1): 15%

2마		예측결과	
		False	True
실제 정답	False	4127	197
	True	717	325

Accuracy: 83%
Recall(1): 31%
Precision(1): 62%
F1-score(1): 42%

3마		예측결과	
		False	True
실제 정답	False	795	691
	True	374	3507

Accuracy: 80%
Recall(1): 53%
Precision(1): 68%
F1-score(1): 60%



Weighted Logistic Regression

1마		예측결과	
		False	True
실제 정답	False	3591	1299
	True	78	398

Accuracy: 74%
Recall(1): 84%
Precision(1): 37%
F1-score(1): 37%

2마		예측결과	
		False	True
실제 정답	False	3175	1210
	True	185	796

Accuracy: 74%
Recall(1): 81%
Precision(1): 40%
F1-score(1): 53%

3마		예측결과	
		False	True
실제 정답	False	1227	313
	True	1066	2760

Accuracy: 74%
Recall(1): 80%
Precision(1): 54%
F1-score(1): 64%

3. 데이터 분석

모델 학습 결과

전체 실험 결과

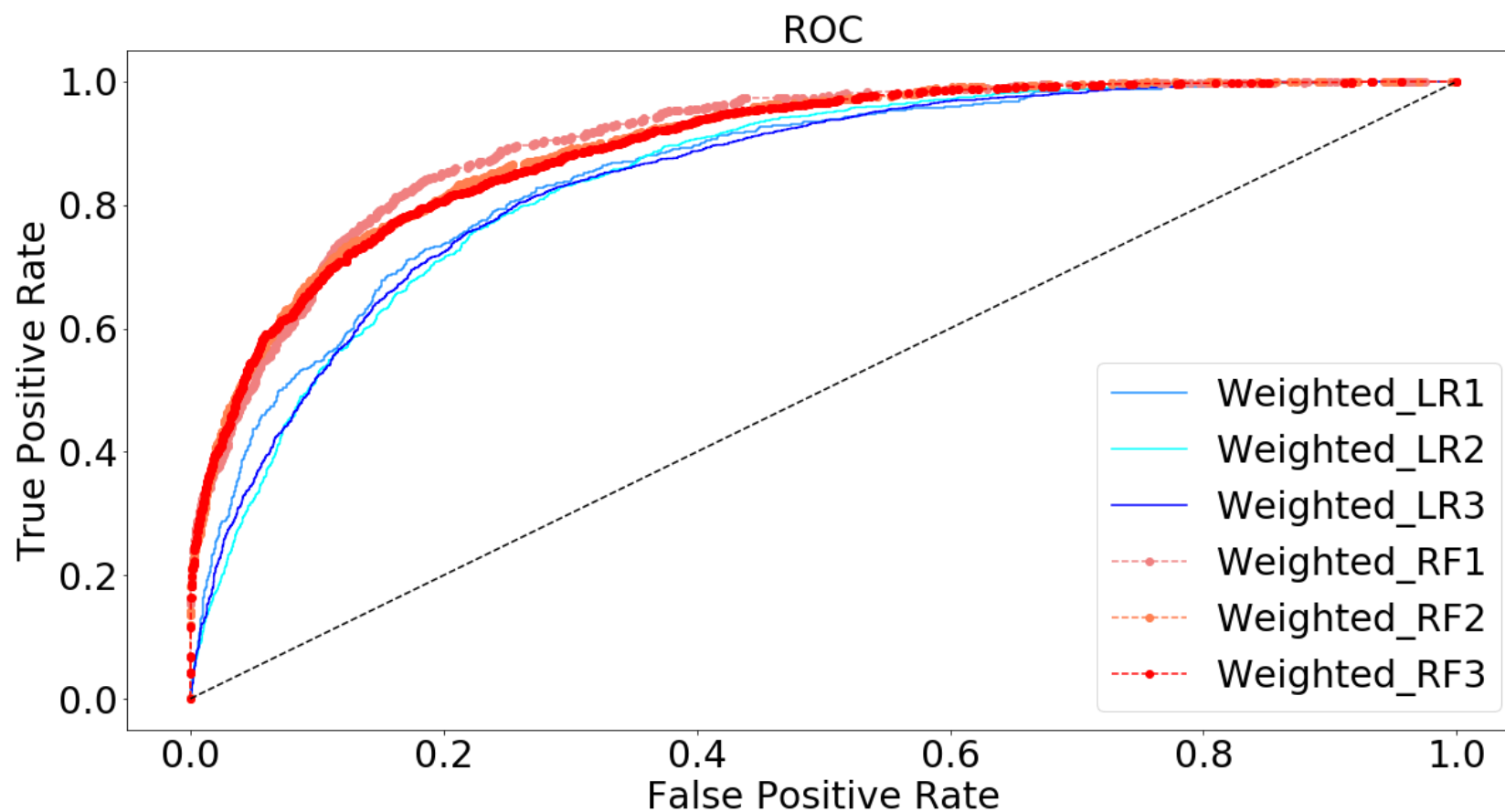
모든 모델에서 가중치를 두고 학습한 결과가 유의미하다고 판단

	Accurcay	Recall(1)	Recall(0)	Precision(1)	Precision(0)	F1-score(1)	F1-score(0)
RF1	93	31.5	99.06	77.83	93.36	44.85	96.13
RF2	88	47.24	96.82	77.38	88.92	58.67	92.7
RF3	84	61.04	93.05	77.26	86.06	68.2	89.42
Weighted_RF1	92	47.5	96.57	58.77	94.7	52.54	95.62
Weighted_RF2	87	59.01	93.72	68.43	90.81	63.37	92.24
Weighted_RF3	84	66.55	90.41	72.88	87.47	69.57	88.92
LR1	91	9	99.26	54.00	92	15	95
LR2	83	31	95.97	62	85	42	90
LR3	80	53	90.55	68	84	60	87
Weighted_LR1	74	84	73	23	98	37	84
Weighted_LR2	74	81	72.46	41	94	54	82
Weighted_LR3	74	80	72.77	54	91	65	81

모델 학습 결과

모델 성능 비교

전체적으로 Random Forest 모델의 성능이 좋다고 판단



결론

정리 및 향후 계획

정리 및 향후 계획

장구현황, 말의 진료사항, 말의 상태 등

유의미한 변수를 찾아 추가하거나, 모델 설정을 다르게 하는 등
다양하고 난이도 높은 분석이 가능함

정확도가 높고, 구체적인 순위를 예측하는 모델 설정도 가능할 것

경마 데이터 분석에 대한 연구가 적은 만큼,
더 깊고 다양한 분석 기법을 활용해 다양한 결과를 도출해보고자 계획 중에 있음

깃허브 주소 : <https://github.com/mjs1995/Horse>