



2018 기상청 빅데이터 콘테스트

날씨에 따른 경마 성적 분석 및 예측

KUBIG팀

주제선정배경

왜 경마인가?

창의성

날씨와의 연관성을 예상할 수 있는 흔한 주제가 아닌

새로운 분야에서의 연구 발상

기상 데이터의 다양한 활용 시도

정확성

일별/시간별 데이터

시간성이 중요한 날씨데이터와 유의미한 통합 가능



경마산업은 국가경제성장을 견인할 수 있는 새로운 **복합문화산업**

2018년 경마 산업 현황

연간 매출액

연 입장객

관련 일자리

경마사업

7조 6459억

경마장

1,300만

2만3천 명

복권사업

3조5천억

야구장

840만

20대 설문 이색데이트 명소로 선정

마필관리사, 조교사 등 경마 관련 일자리 추산

경마는 어렵다?

오늘의 경주

*오늘의 경주는 한국마사회에서 발행하는 경주프로그램 책자를 말합니다.
각 경주에 출주하는 경주마와 기수, 조교사 등의 자료를 볼 수 있고, 과거 전

1 포리스트워드 14회(83.3%) 한국 수목장비 1. 14회(83.3%) 한국 수목장비 2. 14회(83.3%) 한국 수목장비 3. 14회(83.3%) 한국 수목장비 4. 14회(83.3%) 한국 수목장비 5. 14회(83.3%) 한국 수목장비 6. 14회(83.3%) 한국 수목장비 7. 14회(83.3%) 한국 수목장비 8. 14회(83.3%) 한국 수목장비 9. 14회(83.3%) 한국 수목장비 10. 14회(83.3%) 한국 수목장비	김정준 (4) 9.4.4.5 1. 14회(83.3%) 한국 수목장비 2. 14회(83.3%) 한국 수목장비 3. 14회(83.3%) 한국 수목장비 4. 14회(83.3%) 한국 수목장비 5. 14회(83.3%) 한국 수목장비 6. 14회(83.3%) 한국 수목장비 7. 14회(83.3%) 한국 수목장비 8. 14회(83.3%) 한국 수목장비 9. 14회(83.3%) 한국 수목장비 10. 14회(83.3%) 한국 수목장비
2 개리인 14회(83.3%) 한국 수목장비 1. 14회(83.3%) 한국 수목장비 2. 14회(83.3%) 한국 수목장비 3. 14회(83.3%) 한국 수목장비 4. 14회(83.3%) 한국 수목장비 5. 14회(83.3%) 한국 수목장비 6. 14회(83.3%) 한국 수목장비 7. 14회(83.3%) 한국 수목장비 8. 14회(83.3%) 한국 수목장비 9. 14회(83.3%) 한국 수목장비 10. 14회(83.3%) 한국 수목장비	이배 (7) 5.3.4.9 1. 14회(83.3%) 한국 수목장비 2. 14회(83.3%) 한국 수목장비 3. 14회(83.3%) 한국 수목장비 4. 14회(83.3%) 한국 수목장비 5. 14회(83.3%) 한국 수목장비 6. 14회(83.3%) 한국 수목장비 7. 14회(83.3%) 한국 수목장비 8. 14회(83.3%) 한국 수목장비 9. 14회(83.3%) 한국 수목장비 10. 14회(83.3%) 한국 수목장비
3 승승만승 14회(83.3%) 한국 수목장비 1. 14회(83.3%) 한국 수목장비 2. 14회(83.3%) 한국 수목장비 3. 14회(83.3%) 한국 수목장비 4. 14회(83.3%) 한국 수목장비 5. 14회(83.3%) 한국 수목장비 6. 14회(83.3%) 한국 수목장비 7. 14회(83.3%) 한국 수목장비 8. 14회(83.3%) 한국 수목장비 9. 14회(83.3%) 한국 수목장비 10. 14회(83.3%) 한국 수목장비	201억 (4) 9.4.4.5 1. 14회(83.3%) 한국 수목장비 2. 14회(83.3%) 한국 수목장비 3. 14회(83.3%) 한국 수목장비 4. 14회(83.3%) 한국 수목장비 5. 14회(83.3%) 한국 수목장비 6. 14회(83.3%) 한국 수목장비 7. 14회(83.3%) 한국 수목장비 8. 14회(83.3%) 한국 수목장비 9. 14회(83.3%) 한국 수목장비 10. 14회(83.3%) 한국 수목장비

초심자를 위한 경마왕 예상지 보는 방법				
1 동반의강자	36조	03/01 052000 한 2133발행	267배	21
17전 10.4 (82%)	구자선	1볼드매달 51 3-32 장교 8	2-111	22
444(07) 6-36조강양선	기대양자	6행크즈발 511	3F-38.1	23
6899전 670/70120%	기대양자	3행운대발 54 코 배종 2	8D 14.4	24
최범현 11	기대양자	10세리디미 52 1%	3C3 80%	25
2705전 281/27921%	기대양자	7오백매달 541%	6D 13%	26
483K 8	기대양자	11금빛두루 51 3	1 59%	27
3.101만원	기대양자	9디아미아 51 1%	민호 3	28
부Broken Vow / 도Manemaid	기대양자	26-필드 3	483K 8	29
04 08 중동기인성기부복	기대양자	3C 6 38.1 98.7 25	100%	30
05 05 07 비공상기(마미)	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
출주장: 7주	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
1. 양행배전 반도(마미)에 중동기 인성기(마미)	기대양자	26-필드 3	483K 8	29
2. 마미	기대양자	3C 6 38.1 98.7 25	100%	30
3. 대배전자	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
17. 배전배전 배전 배전	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
4. 조교사	기대양자	26-필드 3	483K 8	29
5. 조교사	기대양자	3C 6 38.1 98.7 25	100%	30
6. 기수(기수)	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
7. 기수(기수)	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
8. 나미	기대양자	26-필드 3	483K 8	29
9. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
10. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
11. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
12. 나미	기대양자	26-필드 3	483K 8	29
13. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
14. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
15. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
16. 나미	기대양자	26-필드 3	483K 8	29
17. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
18. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
19. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
20. 나미	기대양자	26-필드 3	483K 8	29
21. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
22. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
23. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
24. 나미	기대양자	26-필드 3	483K 8	29
25. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
26. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
27. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
28. 나미	기대양자	26-필드 3	483K 8	29
29. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
30. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
31. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
32. 나미	기대양자	26-필드 3	483K 8	29
33. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
34. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
35. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
36. 나미	기대양자	26-필드 3	483K 8	29
37. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
38. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
39. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
40. 나미	기대양자	26-필드 3	483K 8	29
41. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
42. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
43. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
44. 나미	기대양자	26-필드 3	483K 8	29
45. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
46. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
47. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
48. 나미	기대양자	26-필드 3	483K 8	29
49. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
50. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
51. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
52. 나미	기대양자	26-필드 3	483K 8	29
53. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
54. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
55. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
56. 나미	기대양자	26-필드 3	483K 8	29
57. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
58. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
59. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
60. 나미	기대양자	26-필드 3	483K 8	29
61. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
62. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
63. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
64. 나미	기대양자	26-필드 3	483K 8	29
65. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
66. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
67. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
68. 나미	기대양자	26-필드 3	483K 8	29
69. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
70. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
71. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
72. 나미	기대양자	26-필드 3	483K 8	29
73. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
74. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
75. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
76. 나미	기대양자	26-필드 3	483K 8	29
77. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
78. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
79. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
80. 나미	기대양자	26-필드 3	483K 8	29
81. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
82. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
83. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
84. 나미	기대양자	26-필드 3	483K 8	29
85. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
86. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
87. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
88. 나미	기대양자	26-필드 3	483K 8	29
89. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
90. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
91. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
92. 나미	기대양자	26-필드 3	483K 8	29
93. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
94. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
95. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
96. 나미	기대양자	26-필드 3	483K 8	29
97. 나미	기대양자	3C 6 38.1 98.7 25	100%	30
98. 나미	기대양자	4C 43.1 19 73 8 925	27-필드 3	31
99. 나미	기대양자	5C 43.1 19 73 8 925	27-필드 3	32
100. 나미	기대양자	26-필드 3	483K 8	29

경마 경주표가 어렵고 복잡해서
어떻게 읽고 경기를 봐야 할 지 모르겠어.
대충 찍어서 걸었다가
우리 팬히 돈만 잃고 오는 거 아닐까?
영화나 보러 가자.

요즘 경마장으로
이색 데이트를 많이 간다고 해!



현재 마사회는 자세하지만,
해석이 어려운 형태로 정보를 제공

국민 대다수가 쉽고 흥미롭게 경마를 즐기기 어려운 구조



경마는 레저 스포츠인가 사행성 도박인가



경마장 찾는 사람 절반은 중독자...국가 차원 대책 필요

여가선용을 모두 경마장에서 소리지르는데 사용하는 사람을 낮춰 "마쟁이" 혹은 "경마꾼"이라고 부르기도 한다.

불법 언더그라운드 경마 도박 규모 75조(형사정책연구원)

경마 중독자의 하루 평균 마권 구입액은 50만원 ...한 달 평균 경마장 방문 횟수 6.5회

경마 예상 방법론은 크게 블러드(혈통 및 마체), 스피드(주파기록) 핸디캐핑, 트립(경주전개), 클래스(승급/강급) 핸디캐핑 등으로 나뉜다.



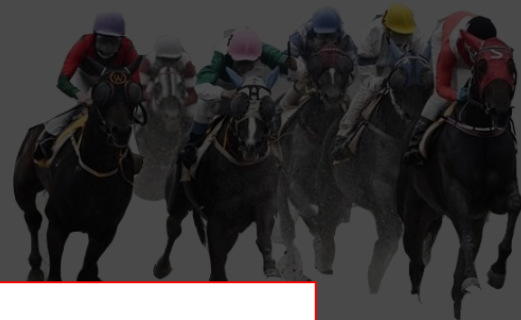
↑↑ 배당! 배당! 배당! ↑↑
합법 토토의 본색파워를 최고배당 확률과 위주로 추천합니다.

7R ③ 성은강자 ~~☆~~
 근성이 부족하여 내측에서 최적

8/18(토) 서울7R **96.3**배

국내 최대 경마 커뮤니티 'ㄱ' 메인 페이지

경마는 레저 스포츠인가 사해성 도박인가



고배당을 강조하는 경마 예측론이 성행

경마가 스포츠가 아닌 **도박**에 가깝다는 **부정적 이미지**

경마장 찾는


여가선용을 모두 경마장에서 소리지르는데 사용하는 사람을 낮춰 "마쟁이" 혹은 "경마꾼"이라고 부르기도 한다.

불법 언더그라운드 경마 도박 규모 75조(형사정책연구원)

경마 중독자의 하루 평균 마권 구입액은 50만원 ...한 달 평균 경마장 방문 횟수 6.5회

경마 예상 방법론은 크게 블러드(혈통 및 마체), 스피드(주파기록) 핸디캐핑, 트립(경주전개), 클래스(승급/강급) 핸디캐핑 등으로 나뉜다.

배당! 배당! 배당!

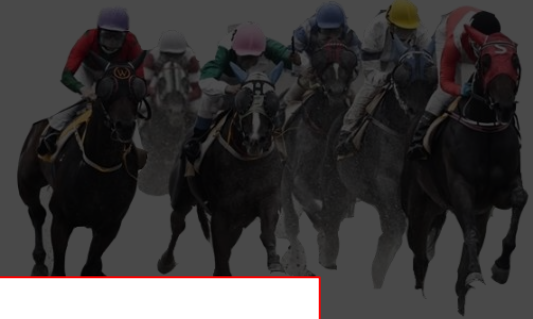
7R ③ 성은강자 

근성이 부족하여 내측에서 최적

8/18(토) 서울7R **96.3**배

국내 최대 경마 커뮤니티 'ㄱ' 메인 페이지

경마는 레저 스포츠인가 사해성 도박인가



고배당을 강조하는 경마 예측론이 성행

경마가 스포츠가 아닌 **도박**에 가깝다는 **부정적 이미지**

여가선용을 모두 경마장에서 소리지르는데 사용하는 사람을 낮춰 "마쟁이" 혹은 "경마꾼"이라고 부르기도 한다.

불법 언더그라운드 경마 도박 규모 75조(형사정책연구원)

경마 중독자의 하루 평균 마권 구입액은 50만원 ...한 달 평균 경마장 방문 횟수 1.5회

경마 예상 방법론은 크게 블러드(혈통 및 트립(경주전개), 클래스(승급/강급) 한다

경마성적 데이터를 **확률적 예측에 활용하는 방법 부재**

배당! 배당! 배당!

7R ③ 성은강자

근성이 부족하여 내측에서 최적

8/18(토) **96.3**배

국내 최대 경마 커뮤니티 'ㄱ' 메인 페이지

그렇다면,
어떻게 이 문제를 해결하여
경마산업이 과학적 분석을 활용해 순조롭게 발전할 수 있을까?

이미 제공되고 있으나 활용되지 못하고 있는 경마데이터를 가공/분석하여

경마 성적과 날씨 정보를 이용한 경주 분석 모델을 제공한다면

이용객들의 정보 활용성이 높아지고, 합리적인 예측이 가능해짐으로써

경마가 사행성 도박이 아닌, 건전한 국민 여가 스포츠로서 기능할 수 있을 것이다.

왜
경마에
날씨 데이터를 이용하는가?

- ✓ 경주 당일의 특성을 반영할 수 있는 일별 변수 부재
- ✓ 날씨 요인에 민감한 동물의 성향 고려
- ✓ 날씨 요인이 육상 기록 단축 스포츠에 미치는 영향 반영

데이터 출처

한국 마사회



출처 : 공공데이터포털에 게시된 한국 마사회 경마정보
기간 : 2015.01.01~2017.12.31

- ✓ 경마장 성적정보
- ✓ 경주 출전표 정보
- ✓ 최근 경주마정보
- ✓ 경주마 훈련현황
- ✓ 경마장 기수정보

기상자료개방포털



출처 : 기상자료개방포털에 게시된 기상청 날씨정보
기간 : 2015.01.01~2017.12.31

- ✓ 종관기상관측자료

데이터 전처리

1. 데이터 병합 (1)

개별 기수, 개별 경주마의 상세정보 데이터와
각 경주별 경마출전표, 경마성적 데이터를
대응시켜 병합한다

기수정보

변수명	변수설명	변수유형
jkName	기수 이름	string
cntT	총 경기수	numeric
ord1T	총 1위 횟수	numeric
ord1Y	최근 1년 1위 횟수	numeric
stDate	데뷔일자	string
...

경주마 정보

변수명	변수설명	변수유형
age	나이	numeric
hrName	경주마 이름	string
cntT	총 경기수	numeric
ord1T	총 1위 횟수	numeric
calt	승군점수	numeric
wgBudam	부담 가능 중량	numeric

출전표 데이터

변수명	변수설명	변수유형
jkName	기수 이름	string
hrName	경주마 이름	numeric
rcTime	경주 시간	numeric
corner	코너별 성적	numeric
differ	앞 순위와 기록차	numeric
rcDate	경주 날짜	string
rcNo	경주 번호	category
wgHr	말 무게	numeric
sex	성별	category
...

데이터 전처리

1. 데이터 병합 (2)

1차 병합된 경마 데이터

변수명	변수설명	변수유형
jkName	기수 이름	string
hrName	경주마 이름	numeric
rcTime	경주 시간	numeric
corner	코너별 성적	numeric
differ	앞 순위와 기록차	numeric
rcDate	경주 날짜	string
rcNo	경주 번호	category
wgHr	말 무게	numeric
sex	성별	category
...

기상 데이터

변수명	변수설명	변수유형
일시	년/월/일/시간	string
기온(°C)	기온	numeric
강수량(mm)	강수량	numeric
풍속(m/s)	풍속	numeric
습도(%)	앞 순위와 기록차	numeric
일사(MJ/m2)	경주 날짜	string
현지기압(hPa)	경주 번호	category
전운량(10분위)	말 무게	numeric
지면온도(°C)	성별	category
...

보다 정확한 날씨 정보를 이용하기 위해
각 경기 별 시행 시간과 시간별 기상정보를 대응시켜 병합한다

데이터 전처리

2. 결측 데이터 제거

세 종류의 경마데이터(출전표+기수+말)를 통합하는 과정에서
결측데이터가 1개의 열이라도 있는 행은 제거한다
36,409개 → 26,539개

date	age	wt	jockey	기온(° C)	강수량(mm)	...
1/3/2017	3	352.3	안효리	3.1	0	...
1/3/2017	8	329.1	페로바치		0	...
1/3/2017	2	376.2		2.8	12	...

3. 불필요한 변수 제거

경주마의 속력을 종속변수로, 기상 요인들은 독립변수로 한 다중회귀분석 결과 유의하지 않다고 판단된 변수들은 제거한다.

기온(° C)	강수량(mm)	습도(%)	평균 이슬점온도 (° C)	해면기압(hPa)	...
3.7	0.4	51	-1	1008	
24.5	83.5	94.6	20.9	1009.2	
31.7	0	30	21	1735	

4. 새로운 변수 정의

경기마다 경주 거리가 달라 단순 피니싱 기록으로는
경주마 별 **공정한 비교**에 어려움이 있을 것으로 판단하여 '속력' 변수를 생성한다.

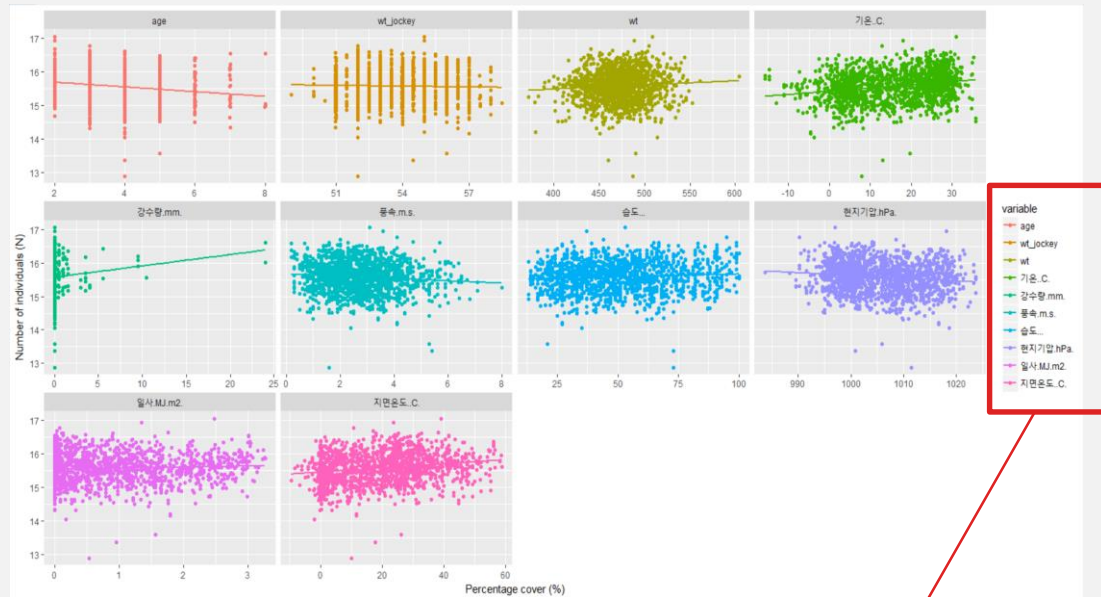
변수명	변수 설명	변수 타입
Y_velocity	경주 거리 / 경주 기록	numeric

Ex) 1200m / 80.2 → 14.9626

hr_name	age	속력
선봉영웅	4	15.1512
비바타운	5	14.72995
선샷	3	16.07177

시각화

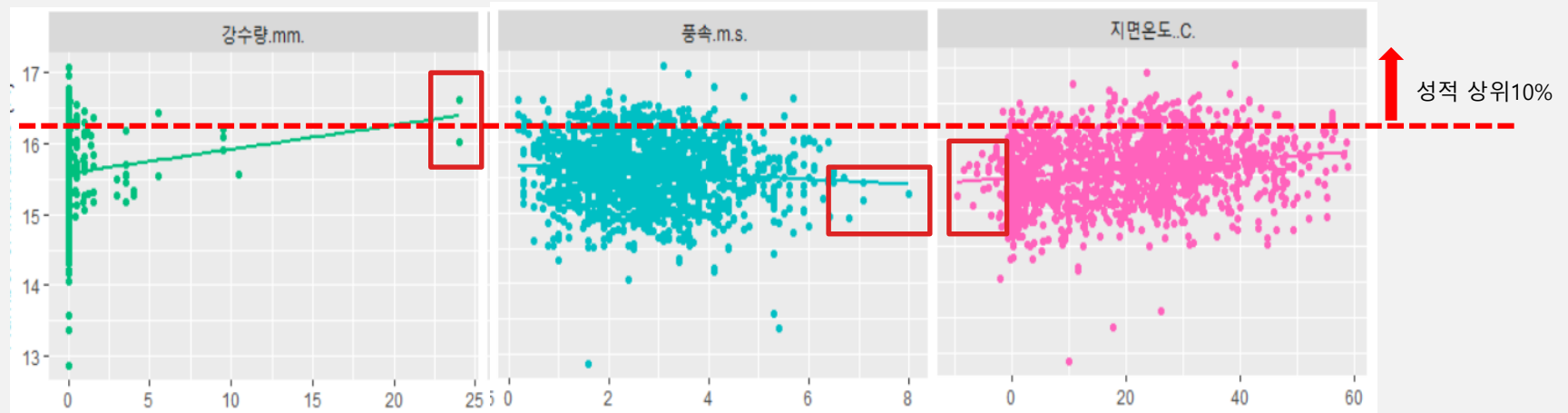
*개별 변수로는 속력과의 연관성을 파악하기 힘들기 때문에 다중회귀를 통한 변수선택 진행



나이/기수체중/말체중/기온/강수량/풍속/습도/현지기압/일사/지면온도

각각의 변수에 대한 산점도와 회귀식

산점도를 이용한 분석 방향 도출



날씨 지표에 대한 속력을 그려본 결과

- ✓ 풍속 및 지면온도가 특정 조건일 경우 상위 성적이 나오지 않는다
- ✓ 강수량이 많아질수록 평균 속력이 상승한다



날씨와 경마 성적 간에 상관성이 있다는 것을 파악했으므로
경마성적 분석에 통계적인 모형을 활용할 것이다.

분석 1.

Stepwise

: 10개의 연속형 변수 중 유의한 변수 선택하기

```
Step: AIC=-45475.49
data1.Y_velocity ~ 기온..C. + age + 습도... + 현지기압.hPa. +
wt + 강수량.mm. + 지면온도..C. + 일사.MJ.m2. + 풍속.m.s. +
wt_jockey
```

	Df	Sum of Sq	RSS	AIC
<none>			4778.6	-45475
- wt_jockey	1	0.600	4779.2	-45474
- 풍속.m.s.	1	0.975	4779.6	-45472
- 기온..C.	1	4.559	4783.2	-45452
- 강수량.mm.	1	5.225	4783.8	-45448
- 일사.MJ.m2.	1	7.026	4785.6	-45439
- 지면온도..C.	1	10.688	4789.3	-45418
- wt	1	11.917	4790.5	-45411
- 현지기압.hPa.	1	16.106	4794.7	-45388
- 습도...	1	40.292	4818.9	-45255
- age	1	101.598	4880.2	-44919

```
Call:
lm(formula = data1.Y_velocity ~ 기온..C. + age + 습도... + 현지기압.hPa. +
wt + 강수량.mm. + 지면온도..C. + 일사.MJ.m2. + 풍속.m.s. +
wt_jockey, data = data2)
```

Coefficients:		기온..C.	age	습도...	현지기압.hPa.	wt	강수량.mm.	지면온도..C.	일사.MJ.m2.	풍속.m.s.
(Intercept)		9.4631214	-0.0585954	0.0025776	0.0058114	0.0008008	0.0105816	0.005950	-0.0405179	-0.0051843
wt_jockey			-0.0027006							

변수명	설명
wt_jockey	기수 체중
wt	말 체중

✓ 10개의 변수 모두 포함하는 것이 최소의 AIC를 갖는다

→ 10개의 연속형 변수와 2개의 범주형 변수(성별, 전문량)을 모두 고려하는 것을 최적의 모형으로 선택

✓ 속력에 음의 상관관계를 보이는 변수 : **나이/일사량/풍속/기수체중** → 속력 저해 요인

분석 2.

다중회귀

```
call:
lm(formula = data1.Y_velocity ~ ., data = data3)

Residuals:
    Min       1Q   Median       3Q      Max
-4.4964 -0.2783  0.0241  0.2934  1.5163

Coefficients: (2 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.1190349   0.6423841  14.196 < 2e-16 ***
age          -0.0588630   0.0025048  -23.500 < 2e-16 ***
wt_jockey    -0.0017383   0.0015338   -1.133  0.2571
wt           0.0009488   0.0001044   9.090 < 2e-16 ***
기온..C.      0.0042724   0.0008690   4.917 8.86e-07 ***
강수량.mm.    0.0094036   0.0019735   4.765 1.90e-06 ***
풍속..m.s.   -0.0052120   0.0022363   -2.331  0.0198 *
습도...      0.0019868   0.0001919  10.355 < 2e-16 ***
현재기압.hPa. 0.0061167   0.0006186   9.888 < 2e-16 ***
일사.MJ.m2.  -0.0486836   0.0067298   -7.234 4.82e-13 ***
지면온도..C.  0.0060805   0.0007769   7.827 5.19e-15 ***
X_0          -0.0661705   0.0093203   -7.100 1.28e-12 ***
X_1           0.0091986   0.0151070    0.609  0.5426
X_2          -0.0637125   0.0159061   -4.006 6.20e-05 ***
X_3           0.0144862   0.0155950    0.929  0.3529
X_4          -0.0043991   0.0159559   -0.276  0.7828
X_5          -0.0544985   0.0172661   -3.156  0.0016 **
X_6          -0.0770146   0.0144210   -5.340 9.35e-08 ***
X_7          -0.0091407   0.0126955   -0.720  0.4715
X_8          -0.0583645   0.0116017   -5.031 4.92e-07 ***
X_9          -0.0231406   0.0115219   -2.008  0.0446 *
X_10         NA         NA         NA         NA
X_1           0.0018316   0.0071697    0.255  0.7984
X_2          -0.0301529   0.0066066   -4.564 5.04e-06 ***
X_3           NA         NA         NA         NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4233 on 26515 degrees of freedom
Multiple R-squared:  0.08182, Adjusted R-squared:  0.08106
F-statistic: 107.4 on 22 and 26515 DF, p-value: < 2.2e-16
```

전운량 0~10분위수
성별
거세말/수암

✓ 전운량/성별 범주형 변수로 더미변수화

→ 모형 적합

✓ 하나의 범주라도 유의하면 모든 범주를 포함시켜야 함

→ 전운량과 성별 유의

✓ 모든 변수를 통제한 상태에서 전운량을 비교

전운량이 1,3분위수일 때: 경주속력 증가

전운량이 4 이상일 때: 경주속력 감소

모든 변수를 통제한 상태에서 성별 비교

거세말이나 암말보다 수말이 가장 속력이 느림

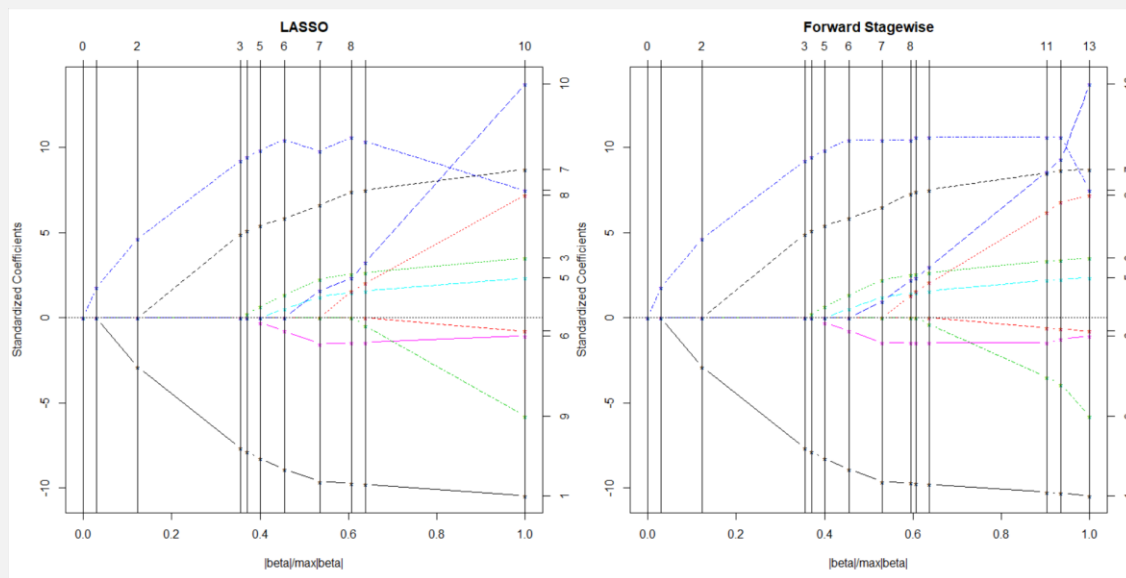
*참고) X_1~X_10은 전운량 1분위부터 10분위수, X_1~X_3은 순서대로 거, 수, 암

분석 3.

Lasso 별점화 함수

: Y에 영향을 가장 크게 미치는 변수는?

Y = 평균속력=거리(m)/시간(초)



```
> coef1
      age      wt_jockey      wt      기온...C      강수량...mm      풍속...m.s.      습도...%      원지...hPa      일사...MJ.m2      지면온도...C
-4.401876e-02  0.000000e+00  5.452122e-05  5.453641e-03  0.000000e+00  0.000000e+00  1.507656e-03  0.000000e+00  0.000000e+00  0.000000e+00
```

✓ 나이/체중/기온/습도

경주 속력에 가장 큰 영향을 미치는 변수

✓ 나이: 음의 상관관계

나이 증가 → 속력 감소

✓ 체중/기온/습도: 양의 상관관계

체중/기온/습도 증가 → 속력 증가

분석 방향

다중회귀

Lasso 벌점화 함수

: 말의 평균 속력에 대한 추정만 가능



단순 평균이 아닌, 속력 상위그룹과 하위그룹에 대한 비교 필요성



Quantile Regression

: 날씨가 각 그룹에 미치는 영향 파악

분석 4.

Quantile Regression

```
Call: rq(formula = data1.Y_velocity ~ ., tau = c(0.1, 0.9), data = data2)
```

```
tau: [1] 0.1
```

하위그룹

```
Coefficients:
```

	Value	Std. Error	t value	Pr(> t)
(Intercept)	11.91960	1.06281	11.21514	0.00000
age	-0.08243	0.00358	-23.00220	0.00000
wt_jockey	0.00021	0.00231	0.09282	0.92605
wt	-0.00047	0.00016	-2.91620	0.00355
기온..C.	0.00496	0.00141	3.51162	0.00045
강수량.mm.	0.02107	0.00044	47.53574	0.00000
풍속.m.s.	0.00740	0.00365	2.02656	0.04272
습도...	0.00247	0.00028	8.74638	0.00000
현지기압.hPa.	0.00329	0.00102	3.21608	0.00130
일사.MJ.m2.	0.02382	0.01071	2.22416	0.02615
지면온도..C.	0.00343	0.00127	2.69920	0.00696

```
Call: rq(formula = data1.Y_velocity ~ ., tau = c(0.1, 0.9), data = data2)
```

```
tau: [1] 0.9
```

상위그룹

```
Coefficients:
```

	Value	Std. Error	t value	Pr(> t)
(Intercept)	7.55925	0.96997	7.79329	0.00000
age	-0.02405	0.00367	-6.54549	0.00000
wt_jockey	0.00800	0.00225	3.55779	0.00037
wt	0.00214	0.00015	14.50557	0.00000
기온..C.	0.00922	0.00126	7.34568	0.00000
강수량.mm.	0.00655	0.00097	6.73360	0.00000
풍속.m.s.	-0.01070	0.00324	-3.29767	0.00098
습도...	0.00305	0.00026	11.51871	0.00000
현지기압.hPa.	0.00689	0.00093	7.38549	0.00000
일사.MJ.m2.	-0.04050	0.00950	-4.26518	0.00002
지면온도..C.	0.00193	0.00111	1.73751	0.08231

- ✓ 다중 회귀분석은 평균에 대한 중요도를 파악할 수 있는 반면, 분위수 회귀로는 각 상위그룹과 하위그룹에게 유의하게 영향을 미치는 변수가 다른 것을 확인

하위그룹의 경우 지면 온도가 중요하나, 상위그룹은 그렇지 않음

기수 체중의 경우 하위그룹에게는 유의하지 않은 요인임에 반해,

상위그룹에게는 유의한 요인

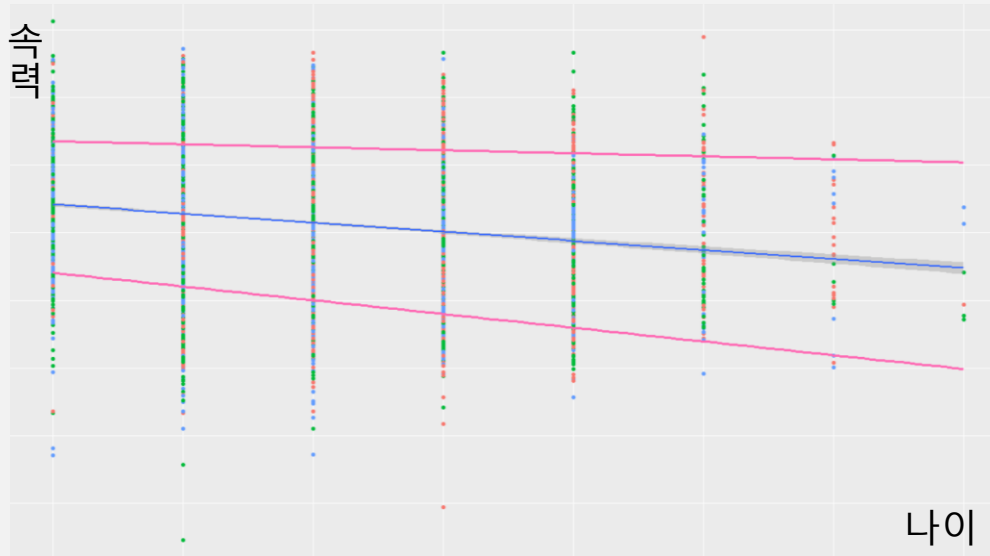
- ✓ 그룹별로 요인의 부호를 살펴보면

상위그룹: 하위그룹과는 다르게 일사량과 풍속에서 음의 영향

하위그룹: 말의 나이와 체중만 속력에 음의 상관관계

분석 4.

Quantile Regression



- 각각의 점은 성별로 **거(주황)**, **수(초록)**, **암(파랑)** 3개
- **위의 분홍색 라인** : 상위 10%그룹의 회귀
- **파란색 선** : 선형회귀
- **아래 분홍색 라인** : 하위 10%그룹의 회귀

- ✓ 성별에 따른 차이는 크게 패턴이 보이지 않음
- ✓ **분홍색 분위수 회귀식**의 기울기를 보면
하위그룹의 경우 나이가 증가함에 따라
속력이 **상위그룹**에 비해 빠르게 감소하는 것을 확인



나이가 2살인 말들과 6세 이상 말들 **각각의 분위수 회귀**를 구해 보고,
과연 나이만으로 인한 차이가 발생하는 것인지
날씨에 영향을 받아서 그런 것인지 확인한다

분석 4.

Quantile Regression

2살 말들의 분위수 회귀

```
Call: rq(formula = Y_velocity ~ ., tau = c(0.1, 0.9), data = data_age2)
```

```
tau: [1] 0.1
```

Coefficients:

	value	Std. Error	t value	Pr(> t)
(Intercept)	9.27879	3.97638	2.33348	0.01969
wt_jockey	0.01694	0.01303	1.29979	0.19378
wt	0.00068	0.00063	1.07393	0.28295
기온..C.	0.00039	0.00570	0.06915	0.94487
강수량.mm.	0.01462	0.00881	1.66000	0.09703
풍속.m.s.	-0.00235	0.01544	-0.15244	0.87885
습도...	0.00181	0.00120	1.50715	0.13189
현재기압.hPa.	0.00457	0.00379	1.20484	0.22837
일사.MJ.m2.	-0.06132	0.03572	-1.71684	0.08612
지면온도..C.	0.00547	0.00482	1.13575	0.25616

```
Call: rq(formula = Y_velocity ~ ., tau = c(0.1, 0.9), data = data_age2)
```

```
tau: [1] 0.9
```

Coefficients:

	value	Std. Error	t value	Pr(> t)
(Intercept)	8.62716	1.97167	4.37557	0.00001
wt_jockey	0.01837	0.00806	2.27855	0.02277
wt	0.00135	0.00040	3.40288	0.00068
기온..C.	0.00935	0.00391	2.39232	0.01681
강수량.mm.	-0.00058	0.01999	-0.02906	0.97682
풍속.m.s.	-0.00657	0.00563	-1.16784	0.24297
습도...	0.00170	0.00052	3.26314	0.00112
현재기압.hPa.	0.00579	0.00191	3.03522	0.00243
일사.MJ.m2.	-0.00509	0.03012	-0.16894	0.86586
지면온도..C.	-0.00392	0.00356	-1.10190	0.27060

날씨에 대한 대부분의 요인이 유의하지 않다

어린 말들의 경우는 날씨에 크게 영향을 받지 않는다

6살 이상 말들의 분위수 회귀

```
Call: rq(formula = Y_velocity ~ ., tau = c(0.1, 0.9), data = data_age6)
```

```
tau: [1] 0.1
```

Coefficients:

	value	Std. Error	t value	Pr(> t)
(Intercept)	6.70159	3.01231	2.22474	0.02624
wt_jockey	0.00965	0.00560	1.72481	0.08476
wt	0.00075	0.00043	1.73854	0.08231
기온..C.	-0.00230	0.00473	-0.48605	0.62700
강수량.mm.	0.04251	0.01940	2.19152	0.02856
풍속.m.s.	0.00529	0.00930	0.56878	0.56959
습도...	0.00155	0.00073	2.13208	0.03316
현재기압.hPa.	0.00697	0.00292	2.38515	0.01719
일사.MJ.m2.	-0.09997	0.03502	-2.85470	0.00436
지면온도..C.	0.01505	0.00472	3.18461	0.00148

```
Call: rq(formula = Y_velocity ~ ., tau = c(0.1, 0.9), data = data_age6)
```

```
tau: [1] 0.9
```

Coefficients:

	value	Std. Error	t value	Pr(> t)
(Intercept)	-2.19041	3.46008	-0.63305	0.52679
wt_jockey	0.05340	0.00746	7.15715	0.00000
wt	0.00279	0.00051	5.48056	0.00000
기온..C.	0.00392	0.00557	0.70302	0.48215
강수량.mm.	0.01952	0.02903	0.67224	0.50153
풍속.m.s.	-0.01214	0.01300	-0.93414	0.35038
습도...	0.00299	0.00089	3.35738	0.00081
현재기압.hPa.	0.01369	0.00339	4.04079	0.00006
일사.MJ.m2.	-0.08465	0.04492	-1.88432	0.05971
지면온도..C.	0.01017	0.00551	1.84654	0.06500

2살 말들에 비해 날씨의 영향을 유의하게 받는다

하위그룹이 나이가 증가할수록 기록이 좋지 않은 것은
단순히 나이 때문이 아니라,

나이가 많은 말들은 날씨에 영향을 많이 받는다는 것을 확인

분석 방향

Quantile Regression

: 그룹별 유의미한 요인 파악



유의미한 변수 간 상호적 관계 파악 필요성

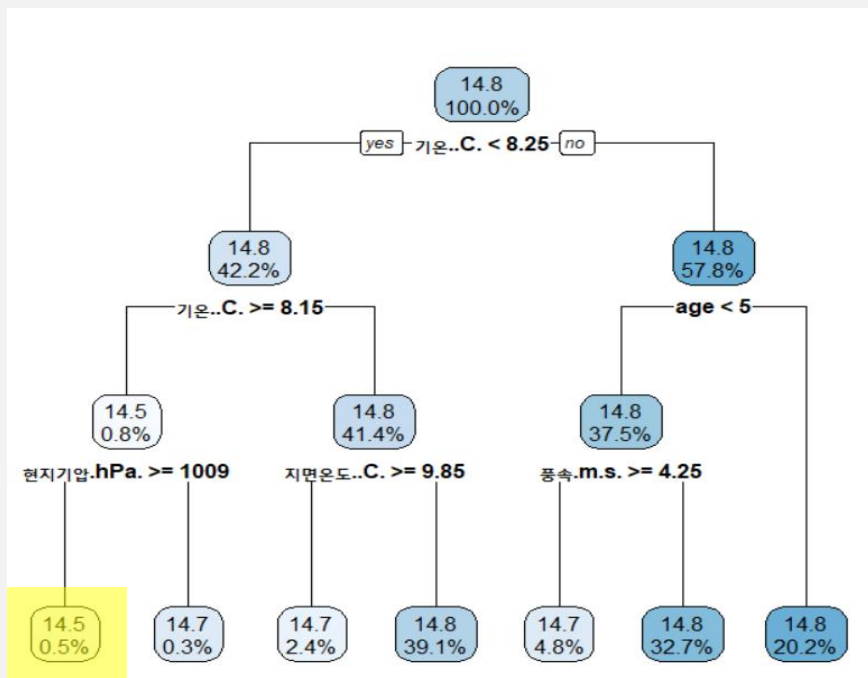


의사결정나무

분석 5.

의사결정나무

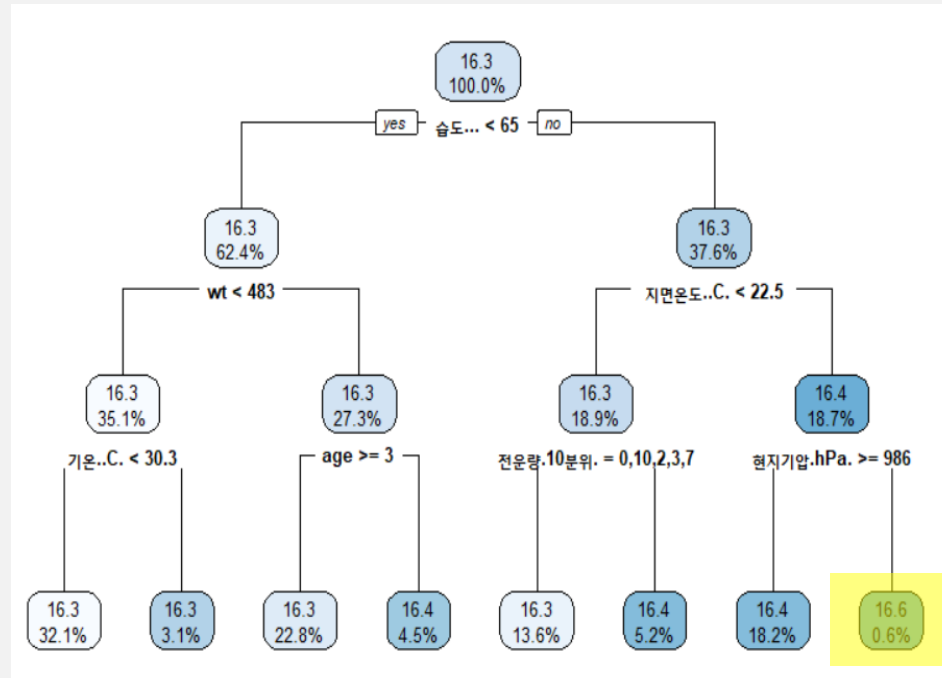
하위10%그룹



- ✓ 기온이 8.15에서 8.25사이일 경우
- ✓ 현지기압이 1009보다 클 경우

평균속력이 가장 낮음

상위10%그룹



- ✓ 습도가 65.5 이상이고
- ✓ 지면온도가 22.45도 이상이며
- ✓ 현지기압이 996 이하일 때

평균속력이 가장 빠름

분석 5.

의사결정나무

하위10%그룹

상위10%그룹

- ✓ 그룹별로 중요한 요인은 특성에 따라 **상호적**으로 판단해야 한다
- ✓ 그룹별로 **날씨**에 따라 말을 관리하고 경기 출전 여부를 판단하거나 조교사의 훈련을 받을 수 있다

- ✓ 기온이 8.15에서 8.25사이일 경우
- ✓ 현지기압이 1009보다 클 경우

평균속력이 가장 낮음

- ✓ 습도가 65.5 이상이고
- ✓ 지면온도가 22.45도 이상이며
- ✓ 현지기압이 996 이하일 때

평균속력이 가장 빠름

분석 방향

의사결정나무 : 변수간 상호적 관계 파악



날씨에 영향을 받는 말 관련 변수들에 대한 가설 수립 및 검증 필요성



패널회귀분석 : 시계열 분석과 횡단면 분석을 동시에 수행하는 회귀분석

분석 6.

패널회귀분석

가설 1.

✓ 말의 특정 **성별**이

날씨에 영향을 받는지

알아보기 위해

패널 분석을 실행한다

Coefficients:

	Estimate	Std. Error	t-value	Pr(> t)
기온..c.	-0.00659642	0.00766862	-0.8602	0.390050
강수량.mm.	0.00071580	0.00470791	0.1520	0.879208
풍속.m.s.	0.00047936	0.00926722	0.0517	0.958765
습도...	-0.00185931	0.00137078	-1.3564	0.175514
현지기압.hPa.	0.00227253	0.00758951	0.2994	0.764721
일사.MJ.m2.	-0.02803385	0.03004579	-0.9330	0.351196
전운량.10분위.	0.00164474	0.00314208	0.5235	0.600861
지면온도..c.	0.00366352	0.00375693	0.9751	0.329907

factor(sex)수	0.03526582	0.01163081	3.0321	0.002539 **
factor(sex)암	0.00431890	0.01186853	0.3639	0.716071
factor(sex)중	-0.16667286	0.16174861	-1.0304	0.303239

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 11.204

Residual Sum of Squares: 10.89

R-Squared: 0.027997

Adj. R-Squared: -0.48358

F-statistic: 1.49254 on 11 and 570 DF, p-value: 0.1299



수컷 말이 날씨의 영향을 더 크게 받는다

분석 6.

패널회귀분석

Coefficients:

	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	9.8155325	5.1952148	1.8893	0.059910 .
기온..C.	0.0087998	0.0069186	1.2719	0.204489
강수량.mm.	0.0083128	0.0190769	0.4357	0.663364
풍속.m.s.	-0.0573703	0.0190429	-3.0127	0.002833 **
습도...	0.0048132	0.0015345	3.1367	0.001896 **
현지기압.hPa.	0.0054301	0.0050843	1.0680	0.286463
일사.MJ.m2.	0.0527275	0.0526509	1.0015	0.317497
전운량.10분위.	-0.0045561	0.0060003	-0.7593	0.448327
지면온도..C.	0.0011673	0.0063124	0.1849	0.853425
factor(name)스트롱샤인	-0.3899724	0.0758182	-5.1435	5.166e-07 ***
factor(name)썬살갈이	0.0711325	0.0729986	0.9744	0.330706
factor(name)아워캣	0.3666048	0.0722171	5.0764	7.142e-07 ***
factor(name)체이스달러즈	-0.0562586	0.0763958	-0.7364	0.462116
factor(name)초원볼트	-0.0014570	0.0747734	-0.0195	0.984468
factor(name)탄성연발	0.1391250	0.0780744	1.7820	0.075872 .
factor(name)햇빛나	0.1455800	0.0744754	1.9547	0.051638 .

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 172.81

Residual Sum of Squares: 26.295

R-Squared: 0.84848

Adj. R-Squared: 0.84013

F-statistic: 101.038 on 15 and 272 DF, p-value: < 2.22e-16

가설2.

✓ 특정 말이

날씨에 영향을 받는지 알아보기 위해,

출전수가 많은 상위 7 마리의 말에 대하여

패널 분석을 실행한다



스트롱샤인, 아워캣

두 말이 날씨에 영향을 잘 받는다

분석 6.

패널회귀분석

Coefficients:

	Estimate	Std. Error	t-value	Pr(> t)
기온..C.	-0.0050598	0.0202819	-0.2495	0.8031
강수량.mm.	0.0267136	0.0299876	0.8908	0.3733
풍속.m.s.	0.0270047	0.0208481	1.2953	0.1956
습도...	0.0020560	0.0036025	0.5707	0.5683
현지기압.hPa.	0.0030769	0.0227564	0.1352	0.8925
일사.MJ.m2.	0.0500543	0.0689073	0.7264	0.4678
전운량.10분위.	0.0086764	0.0070025	1.2390	0.2157
지면온도..C.	0.0024479	0.0079115	0.3094	0.7571
factor(nation)미	-0.6655448	0.0526504	-12.6408	< 2.2e-16 ***
factor(nation)일	-0.0482114	0.0577209	-0.8353	0.4038
factor(nation)캐	0.1153074	0.0732956	1.5732	0.1160
factor(nation)한	-1.0030945	0.0527019	-19.0334	< 2.2e-16 ***
factor(nation)호	-0.2902340	0.0528075	-5.4961	5.139e-08 ***

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 286.95
Residual Sum of Squares: 131.86
R-Squared: 0.54049
Adj. R-Squared: 0.37738
F-statistic: 76.7258 on 13 and 848 DF, p-value: < 2.22e-16

가설3.

✓ 말의 국적 에 따라

날씨에 영향을 잘 받는지 확인하기 위해

5개의 국가의 말에 대해 패널 분석을 실행한다



미국, 한국, 호주의 말이

상대적으로 날씨의 영향을 잘 받는다

데이터 분석을 통한 경마성과 경마, 날씨 요인들의 특성 및 관계 파악



활용 필요성



DNN

: Supervised Learning을 통한 경주 결과 예측

분석 7.

DNN

통계적 분석을 통해, 그룹별로

경주 결과에 영향을 가장 많이 주는 요인들을 선정하여

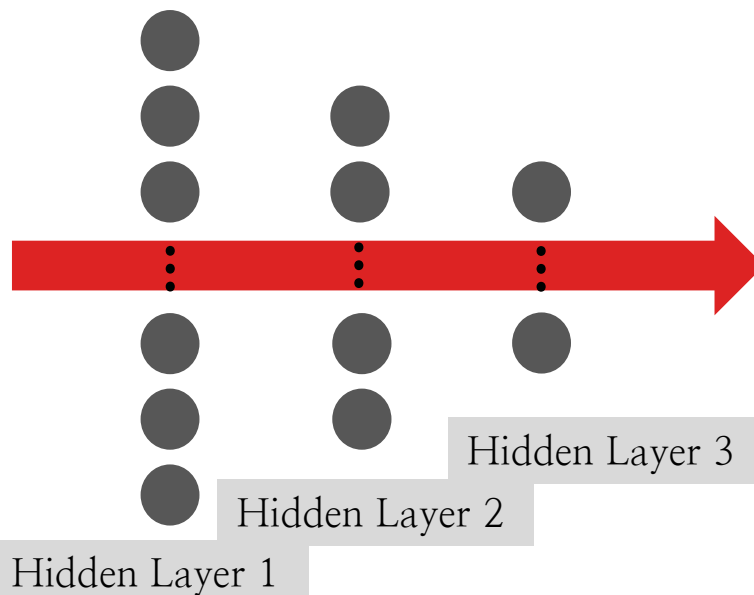
경주 결과 예측 모델을 학습시킨다

Architecture

- ✓ 3 Hidden layers
- ✓ 200, 150, 100 nodes each
- ✓ Initialized with He method
- ✓ ReLU for Activation function

Input

말 무게 (kg)	기수 무게 (kg)	...	기온(° C)	습도(%)
352	51	...	17	54
...
322	56	...	15.6	35



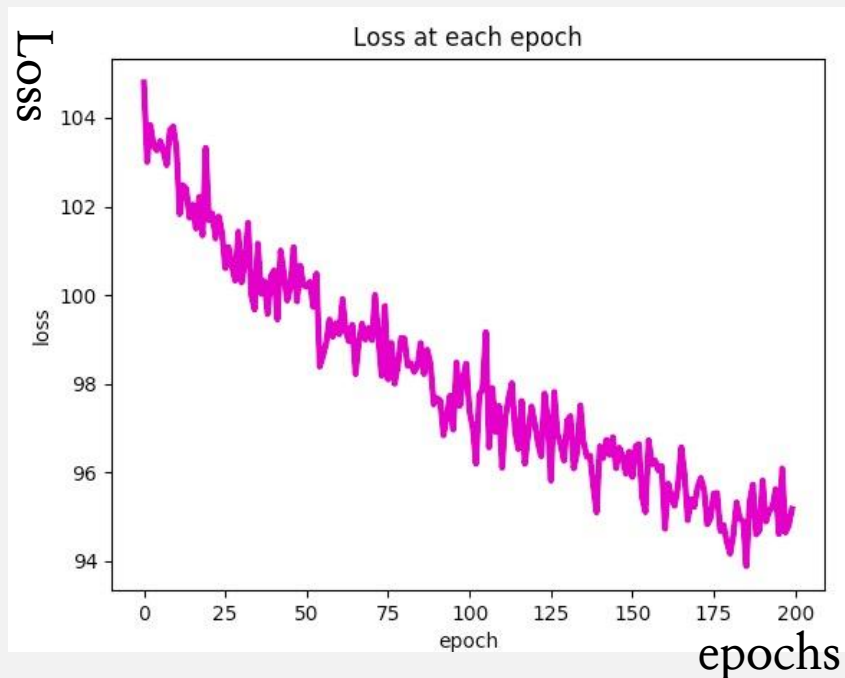
Output

말	예상속력 (km/h)
1번 마	14.86
2번 마	15.27
...	...
...	...
n번 마	14.91

분석 7.

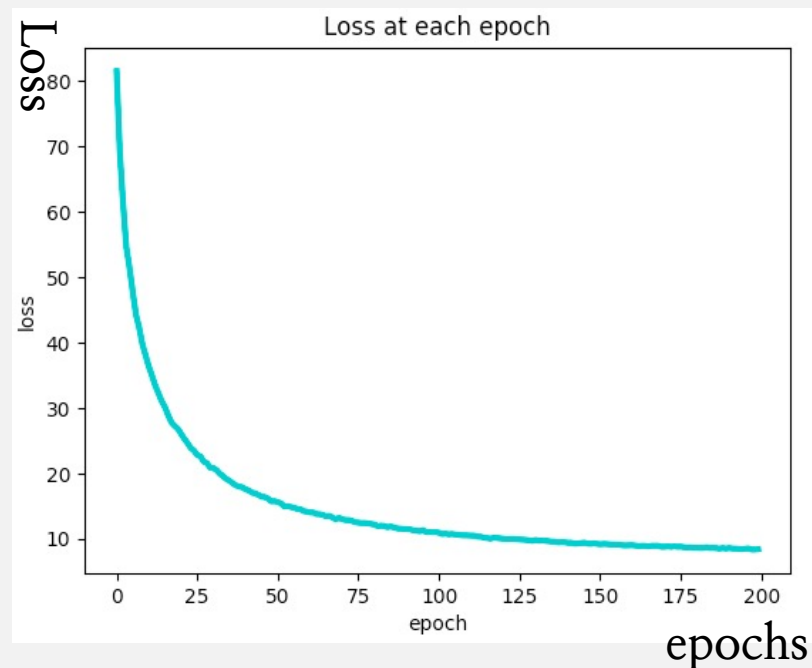
DNN

Learning rate : 0.01



Loss가 explode하는 경향을 보여
적절한 학습이 이루어지지 못함

Learning rate : 0.001

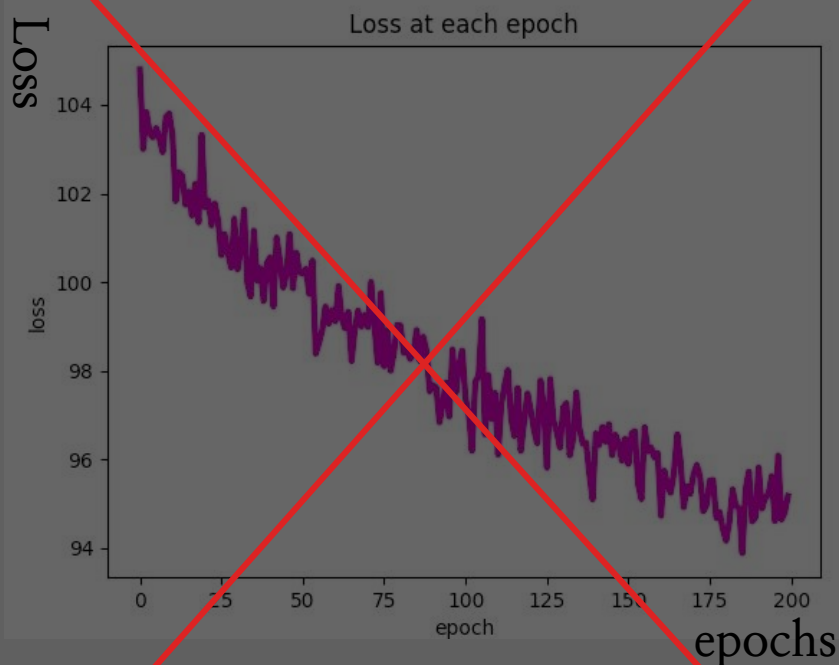


Epoch별 loss가 부드럽게 수렴하며
적절한 예측이 가능하도록 모델이 학습됨

분석 7.

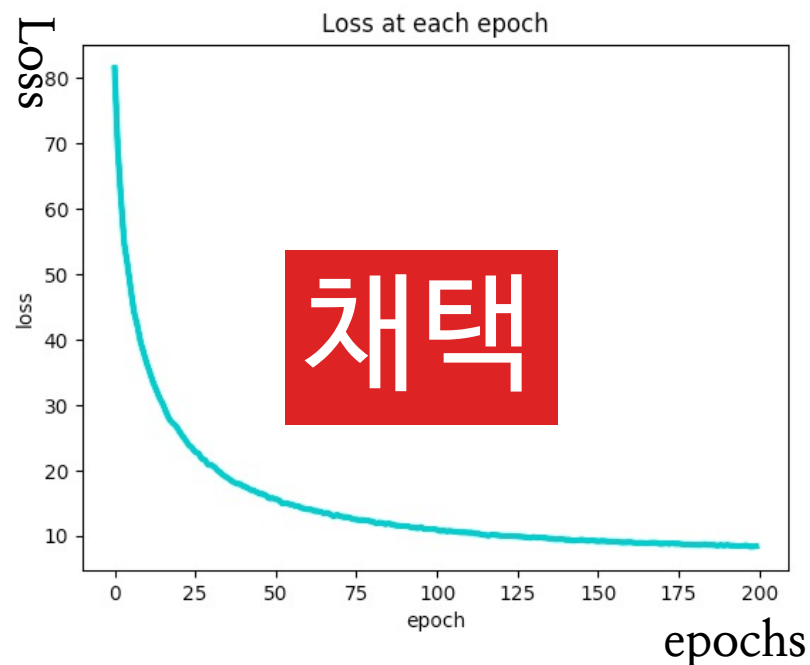
DNN

Learning rate : 0.01



Loss가 explode하는 경향을 보여
적절한 학습이 이루어지지 못함

Learning rate : 0.001



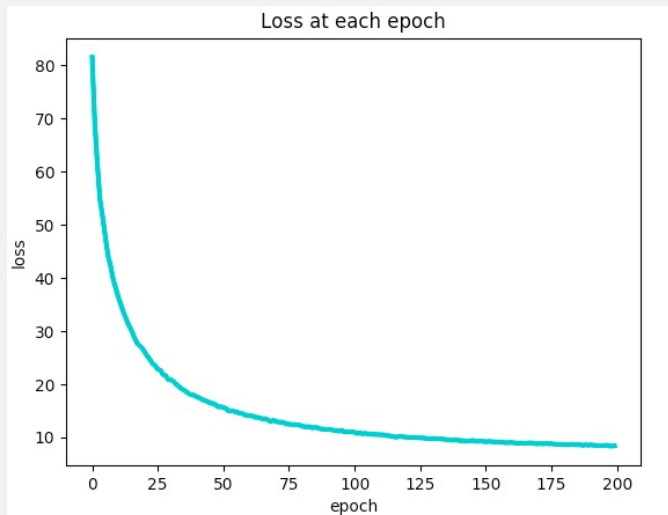
채택

Epoch별 loss가 부드럽게 수렴하며
적절한 예측이 가능하도록 모델이 학습됨

분석 7.

DNN

Learning rate : 0.001



Hyper Parameter

- ✓ 20 epochs training
- ✓ Dropout Rate : 0.5
- ✓ Batch size : 256
- ✓ Learning rate decay : 0.95

Training dataset (2755 games) : 2015.01.01 ~ 2017.06.31
Test dataset (551 games) : 2017.07.01 ~ 2017.12.31

	Top-1마리	Top-2마리	Top-3마리	Top-4마리
DNN	12.4% 50.3%	25.7% 54.8%	36.2% 44.8%	40.8% 22.5%
Random	8.25%	16.6%	25%	33.3%

무작위로 선택하는 도박성 배팅보다

통계적 분석을 통한 합리적 예측의 **승률이 더 높다**는 것을 확인

활용 방안



가족과 연인, 모두 함께 즐기는 경마 스포츠!

이제는
당일의 날씨정보까지 고려한
경마 성적 예측 앱이 있다구!

오 날씨정보까지?
하긴, 우리 동물들한테는
날씨가 정말 중요하니까.
그럼 어디 한 번 해볼까?



우리가 볼 경기는 3경주야.
그때쯤 비가 온다는데, 어떤 말이 관창을까?



조교사님 기수님, 말들 이렇게 훈련시켜 주세요!

조교사와 기수에게 말의 정보를 정확한 데이터 분석을 통해 제공한다.

동물은 **날씨**에 인간보다 더 **민감**하다.

조교사와 기수의 경험에 상위/하위 그룹별, 말 개인별 모형을 더한다.

기온, 습도, 일사량, 전운량 등 어떤 **날씨 요인들**이 각 말들에게 어떻게 영향을 미치는 지를 파악하고 소통하여
경마산업이 과학적으로 한층 더 발전할 수 있다.

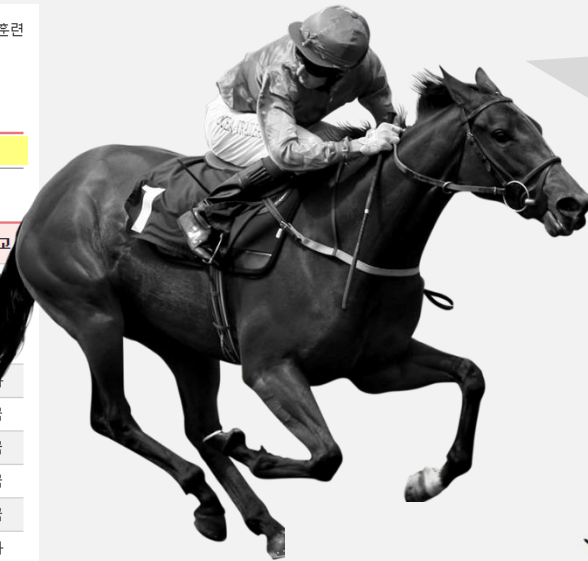
일일훈련

서울경마 > 조교사정보 > 일일훈련

• 일일상세 훈련현황

☑ 마명을 선택하시면 상세정보를 확인하실 수 있습니다.

훈련일자			2017/08/23			소속조		1조(박종곤)			
기승자			일시			평균현지기압			: 조교승인		
순	조		2017-08-21			997.1	입장시각	퇴장시각	훈련시간	걸음걸이	비고
1			2017-08-22			1001.9	08:27	08:39	12분	구보1 습보0	금
2			2017-08-23			1001.3	07:32	07:45	13분	구보0 습보1	금
3							07:11	07:22	11분	구보1 습보0	금
4	10	최강캡틴	외미검	승			06:27	06:43	16분	구보1 습보0	금
5	11	비카스쿠프	국5	조재로			08:41	08:56	15분	구보1 습보0	금
6	13	유니스	국6	승			07:28	07:41	13분	구보1 습보0	금
7	17	테마등극	국1	이혁			07:06	07:18	12분	구보1 습보1	금
8	28	라운메이스	국4	이혁			07:20	07:35	15분	구보0 습보1	금
9	30	수성챔프	국4	승			07:00	07:12	12분	구보0 습보0	차
10	31	라운미라클	국5	승			09:14	09:24	10분	구보1 습보0	차
11	32	라운프로센스	국5	빅투아르			06:37	06:47	10분	구보1 습보0	금
12	33	아레스선더	국5	승			07:47	08:02	15분	구보1 습보0	금
13	38	베스트에버	국6	승			08:18	08:36	18분	구보1 습보0	차



아니,
오늘 이렇게 날씨가 좋은데
라운미라클,
왜
평소보다 기록이 못 나오지?

박종곤 기수님,
저는 기압이 높으면
잘 못 뛰어요ㅠㅠ



일반인에게는,

쉬운 모형 해석을 통해

합리적으로 경기와 배팅,

두 가지를 즐길 수 있게 하고

조교사와 기수등

경마 산업 관련 산업인에게는,

날씨에 따른 분석을 통해

동물을 이해하고

관리할 수 있도록 함으로써

국민들이 직접 날씨 데이터를 주도적으로 활용해 볼 기회를 제공한다.
경마가 과학적인 국민 레저 스포츠로 부상하는 데 기여한다.



감사합니다.

날씨에 따른 경마 성적 분석 및 예측

KUBIG팀

권오준 박소연 이사랑 천우진