


탐색적 데이터분석 미니 프로젝트

이진규



CONTENTS

- 
- 01** **개요**
분석 개요, 가설 수립 및 문헌 조사
 - 02** **데이터 구축 및 분석 방법**
데이터 수집 및 전처리 과정 소개
 - 03** **분석 결과**
분석 결과 확인 및 가설 및 관련 문헌과 비교 후 최종 결론 도출
 - 04** **과제 수행 후기**
문제점 해결 과정 및 느낀 점
 - 05** **마무리**
참고 문헌 및 Q&A

+

01 개요

분석 개요, 가설 수립 및 문헌 조사

1. 분석 배경 소개
2. 분석 목적 소개
3. 분석 범위 설정
4. 가설 수립
5. 관련 문헌 조사



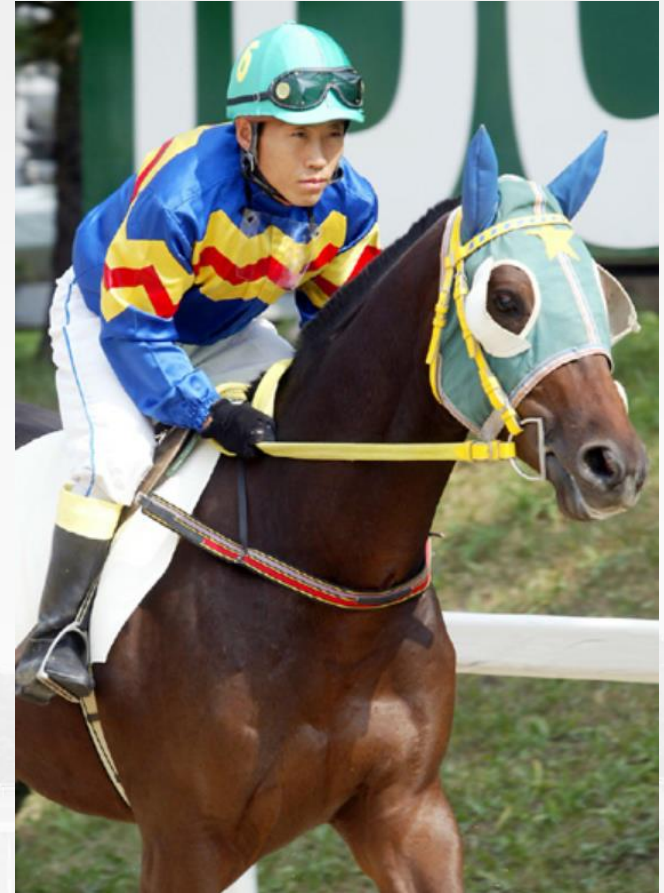
분석 배경 소개

- 분석 대상 선정 : 경마
- 주제 선정 이유
 - 경마는 건전한 스포츠라는 인식과 도박이라는 인식이 공존
 - 고대 그리스, 로마 시대에도 관련 기록이 있을 정도로 유서 깊은 경기
 - 데이터 분석을 통해 우수한 말과 그 성적을 예측 가능한 스포츠인지, 완전한 무작위에 의하여 예측할 수 없는 도박의 영역인지 탐색하고자 함



분석 목적 소개

- 소개에서 설명한 바와 같이 이번 분석의 목적은 경마는 데이터 기반으로 성적이 예측 가능한 스포츠인지, 예측이 불가능한 도박인지 확인하는 것
- 국내의 공식적인 경마 경기는 모두 공기업인 한국마사회에서 주최하고 있으며, 경기 및 출전마에 관련된 정보는 모두 한국마사회에서 공개 중
- 한국마사회 제공 데이터를 기반으로 경마에 관련된 데이터 분석을 진행



분석 범위 설정

- 데이터의 범위는 한국마사회에서 제공하는 국내 경기에 한정하여 분석을 진행
- 보다 정확한 분석을 위해서는 경기의 세부 규정을 확인하여 동등한 기준의 데이터끼리 분석을 진행
 - 말의 레이팅 및 성별에 따른 경주등급 통일
 - 레이팅 : 과거 성적에 따라 말에게 부여하는 등급 점수
 - 분석에 이용할 데이터는 경기를 관전하는 일반인이 사전에 알 수 있는 데이터를 기반으로 선정

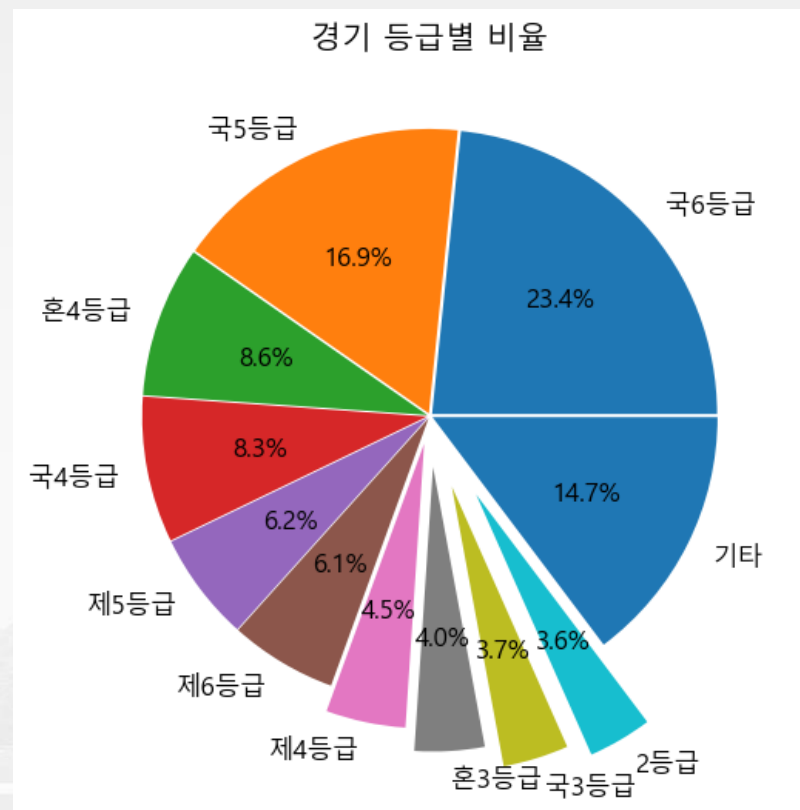


분석 범위 설정

- 경주등급 정보
 - 말의 레이팅 및 연령/성별 분류에 따라 경기군 분리
 - 주로 레이팅의 변화에 따라 상위 티어의 경기로 진출
- 등급분류 규정
 - 데이터가 많은 6등급의 경기로 데이터 분석 진행
 - 6등급 경기는 규정상 국산마만 출전 가능하므로 데이터의 균등한 품질이 유지됨

□ 등급 분류

경주등급	레이팅 구간	연령 · 성별	부담중량	경주거리
대상경주 특별경주	경주 격별 차등 적용	<ul style="list-style-type: none"> • 오픈 • 2세, 3세 • 3세 ↑ • 암 한정 	마령 · 별정	단 ~ 장
1등급	81 이상	<ul style="list-style-type: none"> • 오픈 • 3세 	핸디캡 · 별정	단 ~ 장
2등급	80 이하	<ul style="list-style-type: none"> • 3세 ↑ • 암 한정 	핸디캡 · 별정	단 ~ 장
3등급	65 이하	<ul style="list-style-type: none"> • 오픈 	핸디캡 · 별정	단 ~ 장
4등급	50 이하	<ul style="list-style-type: none"> • 2세, 3세 	핸디캡 · 별정	단 ~ 장
5등급	35 이하	<ul style="list-style-type: none"> • 3세 ↑ ↓ 	핸디캡 · 별정	단 · 중
6등급	미 부 여	<ul style="list-style-type: none"> • 암 한정 	별정	단 · 중



분석 범위 설정

- 일반인이 알 수 있는 사전 데이터
 - (1, 2) : 과거 경기 성적
 - (3, 4, 5, 6) : 과거 경기 관련 기록 (다른 말)
 - (7) : 경기 부담 중량
 - 부담 중량 : 모든 말에게 공정한 경기를 위하여 우수하다고 판단되는 말에게 부여되는 추가 중량
 - (11~) : 기수 정보
- 혈통 정보, 진료 / 훈련 기록 등 추가적인 정보는 한국 마사회에서 정기적으로 제공 중
- 이외 기상 정보, 마번 (부여 코스) 등의 정보를 바탕으로 데이터 분석 진행

1 포리스트월드	김정준 (타) 53.4-0.5	130202 국1 1900 11R 100% 100%	130119 국1 1900 11R 100% 5%	121223 국1 1900 12R 100% 10%	121202 국1 11R 100% 10%
6세(07.03.22) 한국 수목갈매	통산: 52승(59%) 1년: 28승 해당마: 0전(0%)	① 9스퍼블론 2:03.1 57.0 예선(3) ② 10브라더스 2:04.1 54.0 안권(2) ③ 2살배뛰이 2:04.3 51.0 승점(4) ④ 4승리왕 2:04.5 56.5 승점(1) ⑤ 3배작라이 2:04.7 53.0 승점(4) ⑥ 6포리스트 2:05.1 54.0 승점(11) ⑦ 7명고스톱 2:05.5 53.5 승점(5) 15.4-39.2-13.1 / 3C 78.3-4C 93.2 12 12 12 12 11 11 (6) 2 485-4 1617.6 145분 주요(5) 외곽주행	① 3지리이수 2:07.2 56.0 승점(1) ② 9다나리코 2:07.8 59.0 승점(2) ③ 7살그라운 2:07.9 58.0 승점(5) ④ 1살배뛰이 2:07.9 50.0 승점(13) ⑤ 13박토리 2:08.0 53.5 승점(7) ⑥ 8살이스테 2:08.2 56.0 승점(4) ⑦ 4포리스트 2:09.0 54.0 승점(16) 15.1-39.7-13.2 / 3C 81.4-4C 96.9 12 12 12 10 14 13 (13) 4 489-1 165.4 193분 주요(7)	① 13박토리 2:07.3 52.5 승점(9) ② 10살이수 2:07.4 62.0 승점(3) ③ 8살이수 2:07.7 56.0 승점(12) ④ 11포리스트 2:07.7 54.0 승점(7) ⑤ 6살그라운 2:07.8 58.0 승점(5) ⑥ 2살배뛰이 2:07.9 56.5 승점(4) ⑦ 4살리안트 2:09.0 56.0 승점(8) 14.9-39.2-12.7 / 3C 80.6-4C 96.2 14 14 14 14 14 (4) 4 490-2 165.1 133분 주요(7)	① 4스퍼블론 1:58.7 56.0 예선(2) ② 6루비온 1:59.0 55.0 승점(1) ③ 9포리스트 1:59.2 54.0 승점(3) ④ 5살배뛰이 1:59.4 53.0 승점(12) ⑤ 1살배뛰이 1:59.9 52.0 승점(8) ⑥ 7살배뛰이 1:59.9 57.0 승점(10) ⑦ 11살리안트 2:00.1 53.5 승점(4) 15.4-39.4-13.2 / 3C 72.0-4C 87.3 12 13 13 10 11 (6) 3 488-6 150.5 152분 주요(7) 외곽주행(주요)
2 캐러라인	아베 7 52.5-1.5	130126 국2 1900 11R 100% 18%	121222 국2 1900 10R 100% 18%	121224 국2 1900 11R 100% 14%	121028 국2 1400 7R 100% 18%
5세(08.03.08) 한국 양복예 포	통산: 35승(66%) 1년: 15승 해당마: 0전(0%)	① 10캐러라인 2:07.1 54.0 승점(4) ② 2초원여지 2:07.3 53.0 승점(2) ③ 13살이수 2:07.5 56.0 승점(1) ④ 3살배뛰이 2:08.1 57.0 승점(3) ⑤ 11살이수 2:08.5 54.0 승점(7) ⑥ 8살이수 2:08.6 52.5 승점(6) ⑦ 1살배뛰이 2:08.7 52.5 승점(11) 13.7-40.0-13.4 / 3C 78.1-4C 96.0 12 04 17 앞장지 부종4회 120310 앞장지 부종5회	① 12캐러라인 2:06.9 52.0 승점(12) ② 6명수지 2:08.0 51.0 승점(9) ③ 7살배뛰이 2:08.1 52.5 승점(8) ④ 13살이수 2:08.1 50.0 승점(7) ⑤ 5살이수 2:08.1 56.0 승점(2) ⑥ 8살이수 2:08.2 57.0 승점(5) ⑦ 10살이수 2:08.2 53.0 승점(1) 13.9-39.3-13.2 / 3C 79.8-4C 95.2 7 2 6 1 (1) 4 458-2 1675.7 244분 주요(7) (외곽주행)	① 12살이수 1:58.8 55.0 승점(7) ② 9명수지 1:59.2 56.0 승점(9) ③ 8살이수 1:59.5 56.0 승점(1) ④ 4살배뛰이 1:59.6 54.0 승점(2) ⑤ 7캐러라인 1:59.6 52.0 승점(13) ⑥ 10살이수 2:00.1 55.0 승점(11) ⑦ 5살배뛰이 2:00.2 54.0 승점(8) 14.3-39.1-13.1 / 3C 72.6-4C 88.2 5 5 6 7 8 (5) 4 456-8 1623.3 172분 주요(5)	① 5살배뛰이 1:29.3 58.0 승점(3) ② 3살배뛰이 1:29.4 57.0 승점(7) ③ 6캐러라인 1:29.7 54.0 승점(8) ④ 11살이수 1:30.0 55.0 승점(2) ⑤ 1살배뛰이 1:30.3 56.0 승점(5) ⑥ 2살배뛰이 1:30.4 58.0 승점(1) 14.1-40.0-14.2 / 3C 37.4-4C 56.2 11 0 0 10 7 5 (3) 2 448-2 1629.3 275분 주요(3)
3 승승만승	이혁 (타) 53.0-0.4	121117 국1 1900 11R 100% 18% 10	121013 국2 1900 12R 100% 4%	120916 국2 1900 11R 100% 11월달승	120729 국2 1900 10R 100% 18%
5세(08.02.09) 한국 수목갈매	통산: 34승(80%) 1년: 14승 해당마: 0전(0%)	① 12살이수 2:04.3 57.0 승점(7) ② 6포리스트 2:04.4 53.5 승점(2) ③ 11명수지 2:04.5 54.0 승점(1) ④ 2살배뛰이 2:04.5 52.0 승점(5) ⑤ 13살이수 2:05.3 53.5 승점(13) ⑥ 4살배뛰이 2:05.8 53.0 승점(14) 2 주종전: 8회 149(2)번	① 14승승만승 1:57.8 56.5 승점(3) ② 4살배뛰이 1:57.9 56.5 승점(2) ③ 11살이수 1:58.0 56.5 승점(4) ④ 10살이수 1:58.1 56.0 승점(1) ⑤ 6살이수 1:58.4 55.5 승점(5) ⑥ 3살배뛰이 1:59.0 51.5 승점(7) 2 주종전: 8회 149(2)번	① 3살배뛰이 1:57.5 55.0 승점(1) ② 4살배뛰이 1:57.9 55.0 승점(4) ③ 5살배뛰이 1:58.7 58.0 승점(9) ④ 2살배뛰이 1:58.9 58.0 승점(12) ⑤ 11살이수 1:59.2 58.0 승점(3) ⑥ 8살이수 1:59.5 58.0 승점(8) 2 주종전: 8회 149(2)번	① 3살배뛰이 2:05.9 53.0 승점(3) ② 12승승만승 2:06.1 56.0 승점(1) ③ 4살배뛰이 2:07.6 53.0 승점(2) ④ 10살배뛰이 2:07.9 53.0 승점(4) ⑤ 9살배뛰이 2:08.1 53.0 승점(11) ⑥ 7살배뛰이 2:08.4 50.0 승점(12)

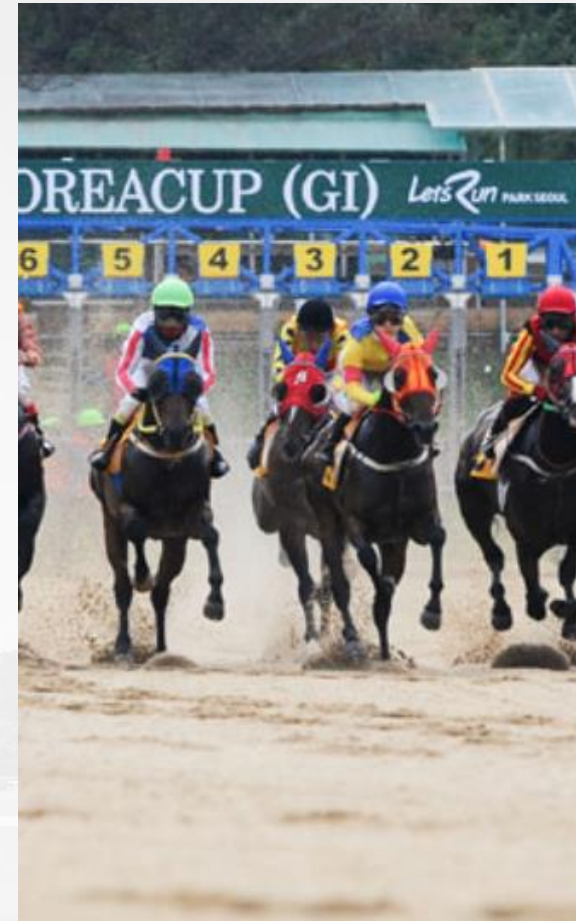
가설 수립

- 잘 달리는 말의 요소를 추측하는 가설 수립

구분	가설
나이	늙은 말일 수록 경주 성적이 떨어진다
성별	숫말이 암말보다 경주 성적이 우수하다
체중	체중은 근육량과 관련이 있으므로 적정 수준 내에서 높은 체중을 보유한 말이 성적이 우수할 것이다

- 경기에 영향을 주는 외부 요소에 관한 가설 수립

구분	가설
날씨	악천후일 경우 전반적인 경기 성적이 하락한다
코스	코스 길이가 길 수록 전반적인 주행 능력이 떨어진다



관련 문헌 조사

- 꾸준한 데이터의 축적 및 돈이 오가는 도박적인 요소에 의한 높은 관심으로 인하여 경마 순위 예측 관련 연구가 상당수 진행됨
- 관전자의 직관적인 노하우는 물론이고, 축적 / 정제된 데이터를 바탕으로 머신러닝이나 딥러닝을 통하여 성적을 예측하려는 시도도 이루어지고 있음
- 이에 따른 다수의 논문이 존재하며 해당 논문들의 가설 및 결론을 이번 데이터 분석과 비교 대조하여 검증하는 과정을 진행
 - 해당 논문 소개 및 검증은 분석 결론에서 다룸



+ 02 데이터 구축 및 분석 방법

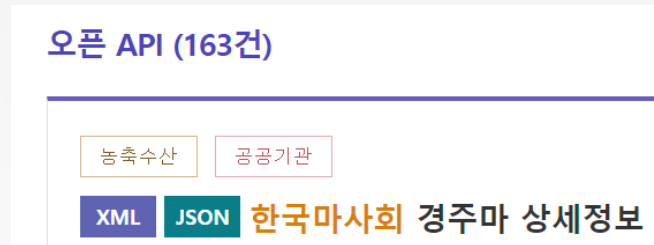
데이터 수집 및 전처리 과정 소개

1. 데이터 수집 과정 및 방법
2. 분석 프로세스 소개
3. 분석 방법 소개



데이터 수집 과정 및 방법

- 국내 경마 경기와 관련된 누적 데이터는 공공데이터 포털에 한국마사회가 정기적으로 업데이트를 진행중
- 오픈 API 기반으로 제공되며 API의 수는 163종류에 달함



- 경기 분석에 필요한 API를 선정하여 최근 5년 (2019 ~ 2023)간의 데이터를 수집하여 데이터 분석에 활용
- 데이터는 JSON 포맷으로 제공되며, 오픈 API이므로 Requests, json 패키지로 모든 데이터 수집이 가능

분석 프로세스 소개



STEP 1

API를 통하여 수집된
데이터 현황 파악

>>

STEP 2

사전에 수립한 분석 기
준에 맞추어 데이터 정
제 진행

1. 이상치, 결측치 처리
2. 분석 대상 필터링
3. 다수 파일은 데이터 병합

>>

STEP 3

Matplotlib, seaborn 의
시각화 라이브러리
등을 이용하여 데이
터 시각화 및 분석 진
행

>>

STEP 4

사전 수립한 가설 및
논문 등의 관련 문헌
과 비교 및 검증 진행

최종 결론 도출

분석 방법 소개

- 데이터 전처리

- 분석의 대상이 되는 변수(종속변수)를 올바르게 찾을 필요가 있음

구분	처리방법
공통	경기후 데이터는 필요한 부분만 사용하고 나머지는 제거 / 결측치, 이상치 확인 및 처리 진행
수치형 데이터	정제된 데이터에 대한 유효성 검증
범주형 데이터	분석에 유효한 카테고리 형식인지, 유효하지 않은 단순 텍스트형 형식인지 확인하여 사용 여부 결정

- 데이터 필터링

- 개요에서 소개한 바와 같이 특정 등급의 경기만을 사용하여 정교한 데이터 분석 진행

- 데이터 분석 및 시각화

- 기초 통계량과 추가적인 통계 기법을 사용하여 데이터 분석 결과를 뒷받침할 합당한 근거 도출
- Matplotlib, Seaborn 을 통한 데이터 시각화 진행, 비교할 데이터에 따라 다양한 유형의 그래프를 활용

분석 방법 소개

수집 대상 데이터

데이터명	내용
경주 정보	경주에 대한 대부분의 데이터 제공 (경주 구분 및 장소, 경주 참여 말/기수, 기상/트랙 상태, 경기 결과 등)
경주마 정보	현역 및 은퇴한 말에 대한 각종 정보 제공 (부모 말, 통산 상금 및 상위권 기록, 연령 등)
마필진료정보	건강 이상으로 이하여 진료받은 내역 제공 (날짜, 대상 말, 병명 등)
마필훈련정보	말이 훈련 받은 내역 제공 (날짜, 구보 및 습보 훈련 횟수, 훈련 시간 등)

데이터 병합 방법

- 모든 데이터는 대상 말에 대한 고유 마번(hrNo)이 부여되어 있어 해당 컬럼을 기준으로 데이터 병합(merge)이 간편하게 이루어짐

hrName	hrNo	hrT
패스트건	46029	계륜
다이아보스	45215	Tria
황우머니	45403	계륜
케이골드바	45845	Tria
캡틴코만도	45167	계륜
해피블루	46394	계륜

hrName	hrNo	meet
소우굿	51581	부산경남
광해대부	3103255	제주
돌체	3103257	제주
타이밍	3103260	제주
줄라이스타	47562	서울
계림비각	3103271	제주
하이폴리	47548	서울
두력산	3103274	제주

분석 방법 소개

- 데이터 파악 – 경주 정보 : 경기에 대한 정보 및 결과에 대한 필수적인 데이터가 포함된 분석의 핵심이 되는 데이터, 모든 데이터 취합은 이 데이터에 가산하는 방식으로 진행
- 경주 정보 데이터는 122,769행, 90열 구성

```
1 df.shape
✓ 0.0s
(122769, 90)
```

- 주요 데이터 목록

컬럼명	타입	내용	컬럼명	타입	내용
age	정수형	말의 나이	rcTime	실수형	말이 완주하는데 걸린 시간
birthday	정수형	말의 생일	sex	텍스트	성별 (수컷 암컷 거세)
budam	텍스트	부담 중량을 결정하는 적용 기준	track	텍스트	주로 상태 및 습도 (건조, 보통 등)
ord	정수형	최종 순위	weather	텍스트	날씨 상태 (맑음, 비 등)
rank	텍스트	경기 등급	wgBudam	실수형	부담 중량
rcDate	정수형	경기 날짜	wgHr	텍스트	마체중

분석 방법 소개

- 데이터 파악 – 기타 파일
 - 각 파일에 대한 간략한 정보 및 사용할 컬럼 선정
 - 경주마 정보 : 부모 정보, 통산 및 최근 1년간 상위권 기록 횟수, 통산 상금
 - 마필진료정보 : 진료일자, 진료 병명
 - 마필훈련정보 : 훈련일자, 구보 및 습보 훈련 횟수
 - 각 데이터의 크기 파악

```
1 경주마정보.shape, 진료정보.shape, 훈련정보.shape
[6] ✓ 0.0s
... ((56445, 26), (207804, 8), (2020082, 15))
```

분석 방법 소개

- 분석 대상 변수 (종속변수) 선정
 - 파생변수 'hrSpd'(말의 속도)으로 선정 = 주행거리(rcDist) / 주행시간(rcTime)으로 계산
 - 단위는 (m/s)

- hrSpd 파생변수 생성

```
1 df2['hrSpd'] = df2['rcDist'] / df2['rcTime']
```

- 순위(ord)가 아닌 이유?
 - 순위는 단일 경기 내의 성적이기 때문에 모든 경기 기록끼리 대조하는 분석에는 맞지 않음
 - 1착을 기록한 말이 정말로 성적이 우수한 말인지, 혹은 대진운이 좋아서 1착을 했는지 알 수 없음
 - 말의 속력을 분석 기준으로 하여도 순위 예측이 가능함 (내림차순 정렬)
 - 데이터 분석의 시각화를 간단하게 할 때 순위를 종속변수로 설정할 수는 있음

분석 방법 소개

- 데이터 전처리 - 데이터 타입 변경
 - 정수 형태로 저장된 날짜 데이터 : birthday(생일), rcDate(경기날짜)
 - Datetime64 타입으로 변환하여 날짜 기반 연산에 사용할 수 있도록 가공

birthday	birthday
20200416	2020-04-16
20200210	2020-04-16

- 텍스트 형태로 저장된 숫자 데이터 : wgHr (마체중)
 - Int64 타입으로 변환하여 데이터 분석에 사용할 수 있도록 가공

wgHr	wgHr
496(-8)	496
461(-5)	495

분석 방법 소개

- 데이터 전처리 - 데이터 필터링

- 개요에서 설명한 바와 같이 rank (경기 등급)은 '국6등급' 으로 제한하여 데이터 필터링 진행

```
1 df2 = df[df['rank'] == '국6등급'].copy()
```

- 데이터 건수는 28,031건으로 감소

- 데이터 전처리 - 데이터 단순화

- 트랙의 상태는 어느정도 카테고리화 되어 습도부분은 불필요한 것으로 판단하여 제거



track
포화 (16%)
포화 (16%)
양호 (9%)

track
포화
포화
양호

- 데이터 전처리 - 파생변수 생성

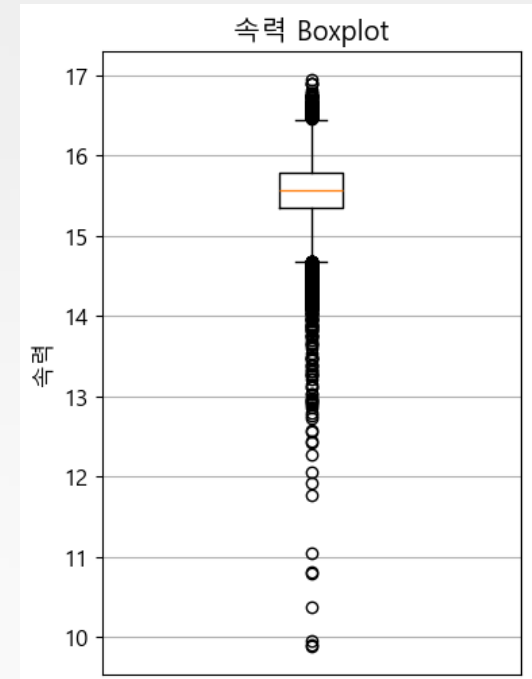
- 연단위 age로는 완전한 데이터 분석이 불가하다고 판단하여 생일과 경기일자의 차이를 이용하여 일단위 나이 계산

```
df2['age_day'] = (df2['rcDate'] - df2['birthday']).apply(lambda x : x.days)
```


분석 방법 소개

- 데이터 전처리 - 이상치

- 속력 Boxplot을 이용하여 데이터의 분포 파악
- 이상치 범위를 벗어나는 데이터의 처리
 - 결론 : 별도로 처리하지 않음
 - 이유 : 등재된 데이터 자체가 경기가 정상적으로 진행되어 올바르게 기록된 데이터이며 주관적인 판단에 의하여 가능하다고 판단되는 범주의 속력이기 때문에 Boxplot 상에서 이상치로 기록되었다고 하더라도 처리를 할 근거가 될 수 없음
- ord에 대한 '규정상' 이상치가 존재
 - 1 ~ 16위까지의 정상적인 데이터 외에 실격, 경기취소 등의 비정상적인 경기 데이터는 ord가 90번대로 등재됨
 - 해당 데이터는 삭제하여 이상치 처리



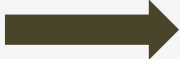
ord	ordBigo
92	주행중지
95	출전취소
95	출전취소
91	실격
92	주행중지
94	경주제외
99	주행중지

분석 방법 소개

• 데이터 전처리 - 결측치

- 결측치 확인 결과 : 날씨(weather) 및 연령대(ageCond)에서 결측치 확인
 - ageCond : 나이(age) 데이터를 참조하여 적합한 카테고리로 배정

age	ageCond
3	연령오픈
3	연령오픈
3	
3	
2	



age	ageCond
3	연령오픈
3	연령오픈
3	3세
3	3세
2	2세

```
1 df2.isnull().sum().so
✓ 0.1s
ageCond      67
weather      33
age           0
```

➤ weather : 인근값 대체 옵션

- 시계열 데이터에 한정하여 데이터가 시간순으로 나열되어 있을 때 결측치 처리는 인근값으로 대체 가능 (ffill : 앞의 데이터, bfill : 뒤의 데이터, .interpolate() : 선형비례 보간 (수치형 한정))

```
1 df2['weather'] = df2['weather'].fillna(method = 'ffill')
```

track	weather
양호 (6%)	맑음
다습 (10%)	
양호 (6%)	맑음
다습 (10%)	맑음
양호 (6%)	맑음
다습 (10%)	맑음
다습 (10%)	



track	weather
양호 (6%)	맑음
다습 (10%)	맑음
양호 (6%)	맑음
다습 (10%)	맑음
양호 (6%)	맑음
다습 (10%)	맑음
다습 (10%)	맑음

분석 방법 소개

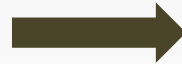
- 데이터 병합

- 훈련데이터 / 진료기록 데이터의 병합

- 훈련데이터는 훈련기록이 2주까지 유효하다는 가정하에 데이터에 훈련 기록의 유효 마감일을 설정하여 경기 데이터에 추가함

```
df_tr['trDate'] = df_tr['trDate'].apply(lambda x : pd.to_datetime(str(x)))  
df_tr['dateEnd'] = df_tr['trDate'] + dt.timedelta(days = 14)
```

trDate	trName	trTerm	dateEnd
2023-01-30	박종곤	660.0	2023-02-13



rcDate	rcNo
2023-01-29	3
2023-02-10	4
2023-03-19	4

- 진료데이터는 진료기록 전후 1주일간을 유효기간으로 설정하고 훈련데이터와 동일한 방법으로 경기 데이터에 추가

```
df_clinic['min_date'] = df_clinic['date'] - dt.timedelta(days = 7)  
df_clinic['max_date'] = df_clinic['date'] + dt.timedelta(days = 7)
```

분석 방법 소개

- 데이터 병합
 - 말 고유 번호 (hrNo)를 기준으로 데이터 병합 가능
 - hrNo만 매칭하면 될 경우 pd.merge로 데이터 병합
 - Inner join을 사용했음에도 경기 데이터의 손실이 없었음
 - hrNo가 누락없이 모두 기록됨

```
1 df2.shape
```

✓ 0.0s

```
(119892, 28)
```

```
1 df2 = pd.merge(left = df2, right = df_hr[['hrNo', 'topCntY', 'topCntT']], how = 'inner', on = 'hrNo')
```

✓ 0.0s

```
1 df2.shape
```

✓ 0.0s

```
(119892, 30)
```


분석 방법 소개

- 전처리 후 최종 데이터 컬럼 리스트 및 설명

컬럼명	타입	내용	컬럼명	타입	내용
age	정수형	말의 나이	rcNo	정수형	마번 (부여 경주로)
ageCond	텍스트	말의 나이대 (마사회 규정에 따름)	rcTime	실수형	완주 소요시간
birthday	날짜	말의 생일	rcDist	정수형	경기 거리
budam	텍스트	경기에 적용된 부담 중량 규정	sex	텍스트	성별 (수컷, 암컷, 거세마)
trackcat	텍스트	주로 상태	wgBudam	실수형	부담 중량
weather	텍스트	날씨 상태	wgHr	정수형	마체중
hrNo	정수형	말 고유번호	hrSpd	실수형	말의 속력 (종속변수)
ord	정수형	경기 순위	age_day	정수형	말의 나이 (일단위)
rank	텍스트	경기 등급	tr1	정수형	경기 전 2주일간 구보 훈련횟수
rcDate	날짜	경기 날짜	tr2	정수형	경기 전 2주일간 습보 훈련횟수
topCntY	정수형	최근 1년간 1~3착 횟수	ills	정수형	경기 전후 1주일간 진료 횟수
topCntT	정수형	통산 1~3착 횟수	meet	텍스트	경기 장소

03 분석 결과

분석 결과 도출 및 가설 및 관련 문헌과 비교

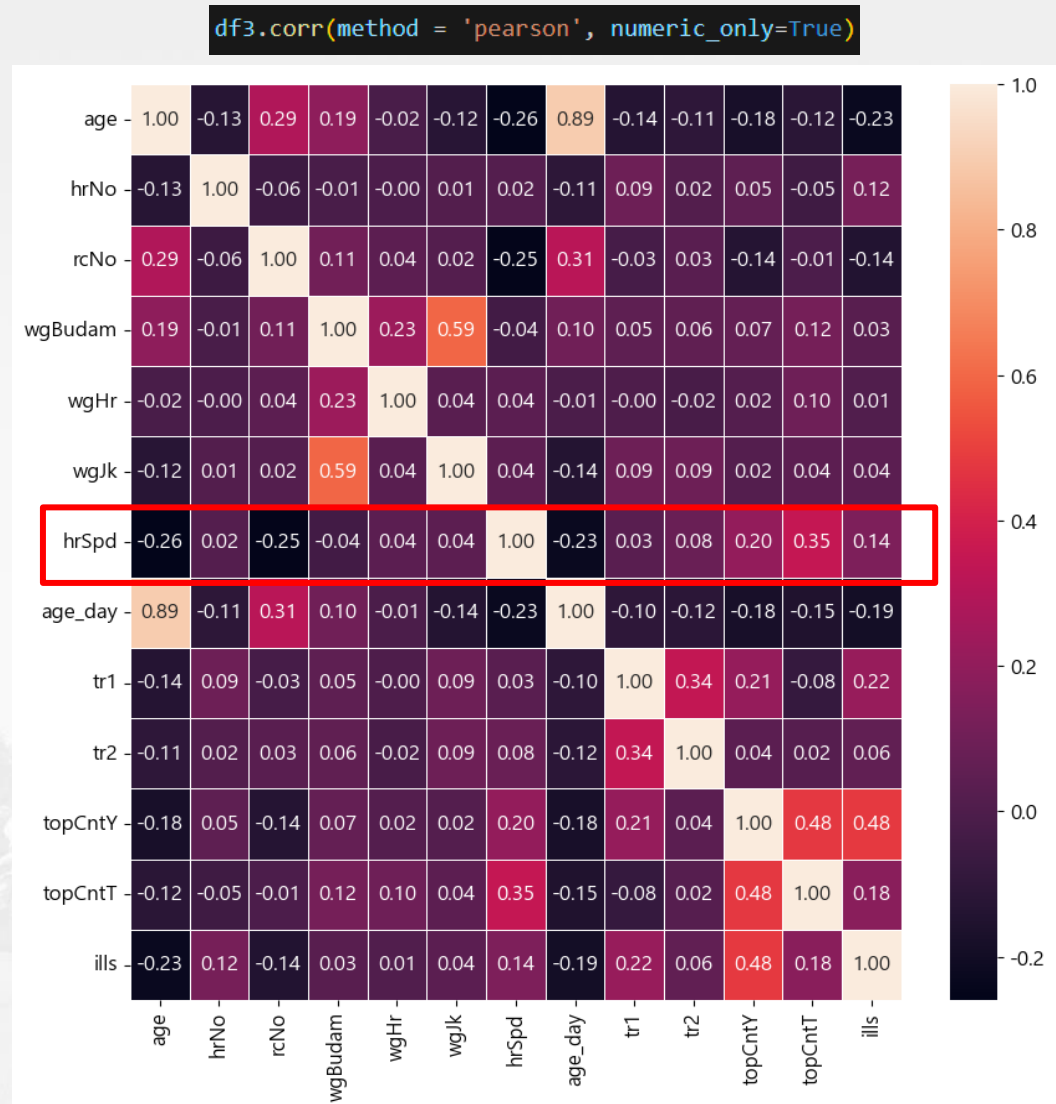
1. 분석 결과 도출
2. 가설 검증
3. 문헌 검증
4. 결론



분석 결과 도출

상관계수 시각화

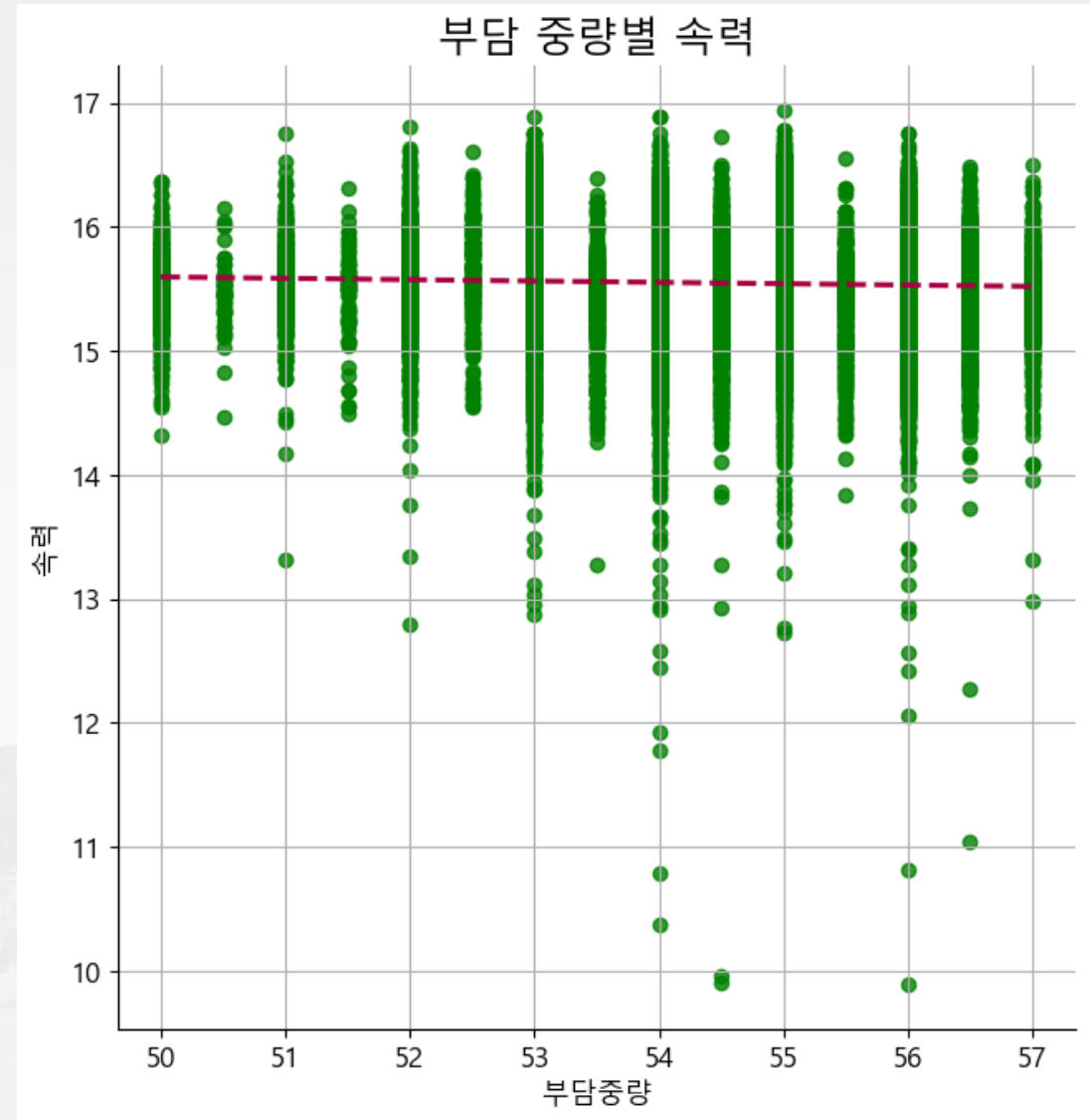
- 피어슨 상관계수 : `df.corr(method = 'pearson')(default))`
- 변수간 어떤 선형 관계가 있는지 수치적으로 파악 가능
- 1 ~ 1의 수치를 가지며 절댓값이 1에 가까울수록 강한 상관관계, 0에 가까울수록 약한 상관관계, 양수이면 양의 상관관계 음수이면 음의 상관관계
- seaborn의 heatmap을 사용하여 시각화 진행
- 종속변수 hrSpd와 상관계수가 큰 값을 파악



분석 결과 도출

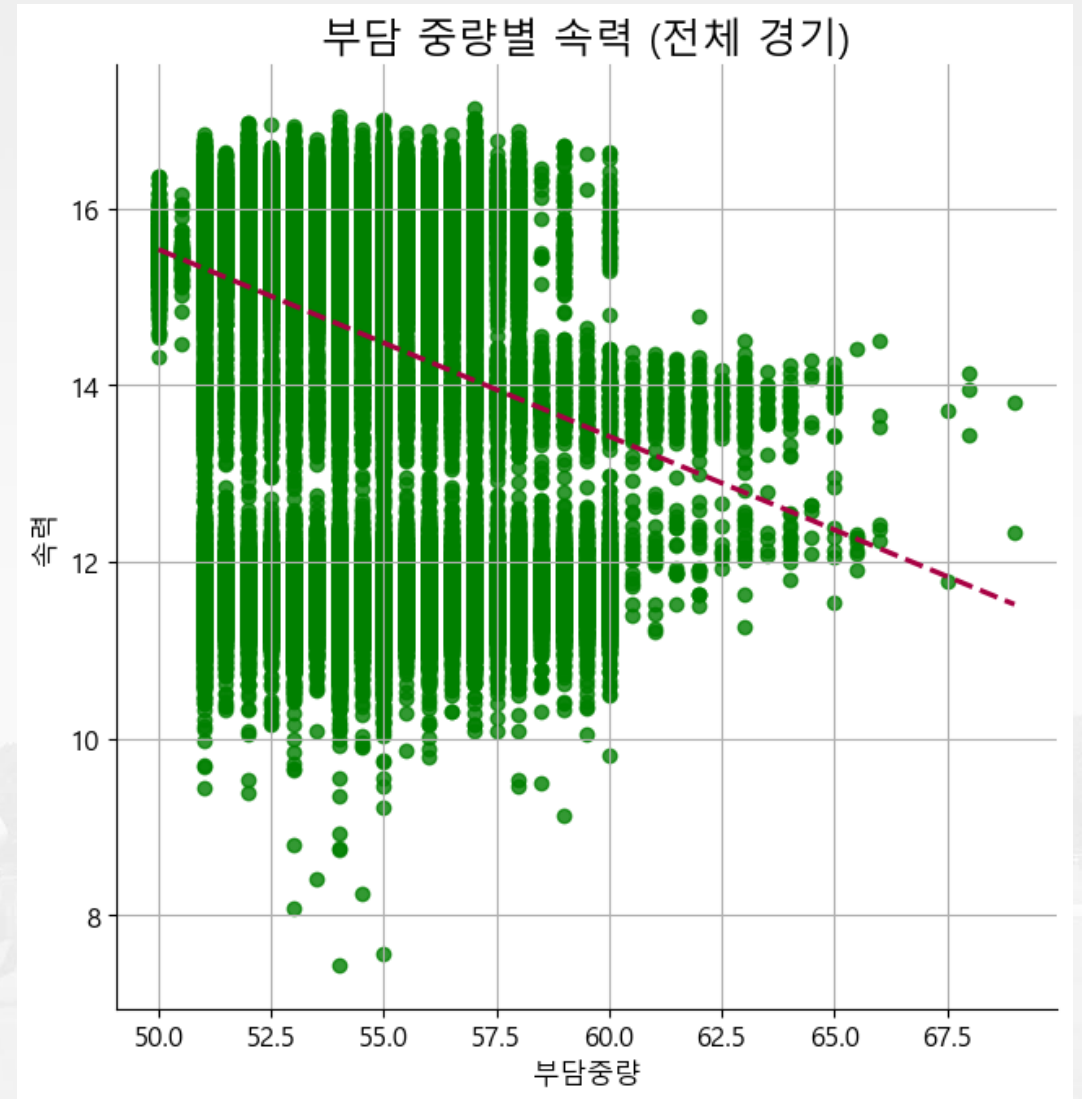
- hrSpd와의 상관관계 시각화 – 부담중량
 - 피어슨 상관계수 = -0.04
 - Implot 사용
 - 상관계수로나 그래프로나 상관관계가 거의 없는 것으로 파악됨

➤ 부담중량으로는 말의 성적을 예측할 수 없음



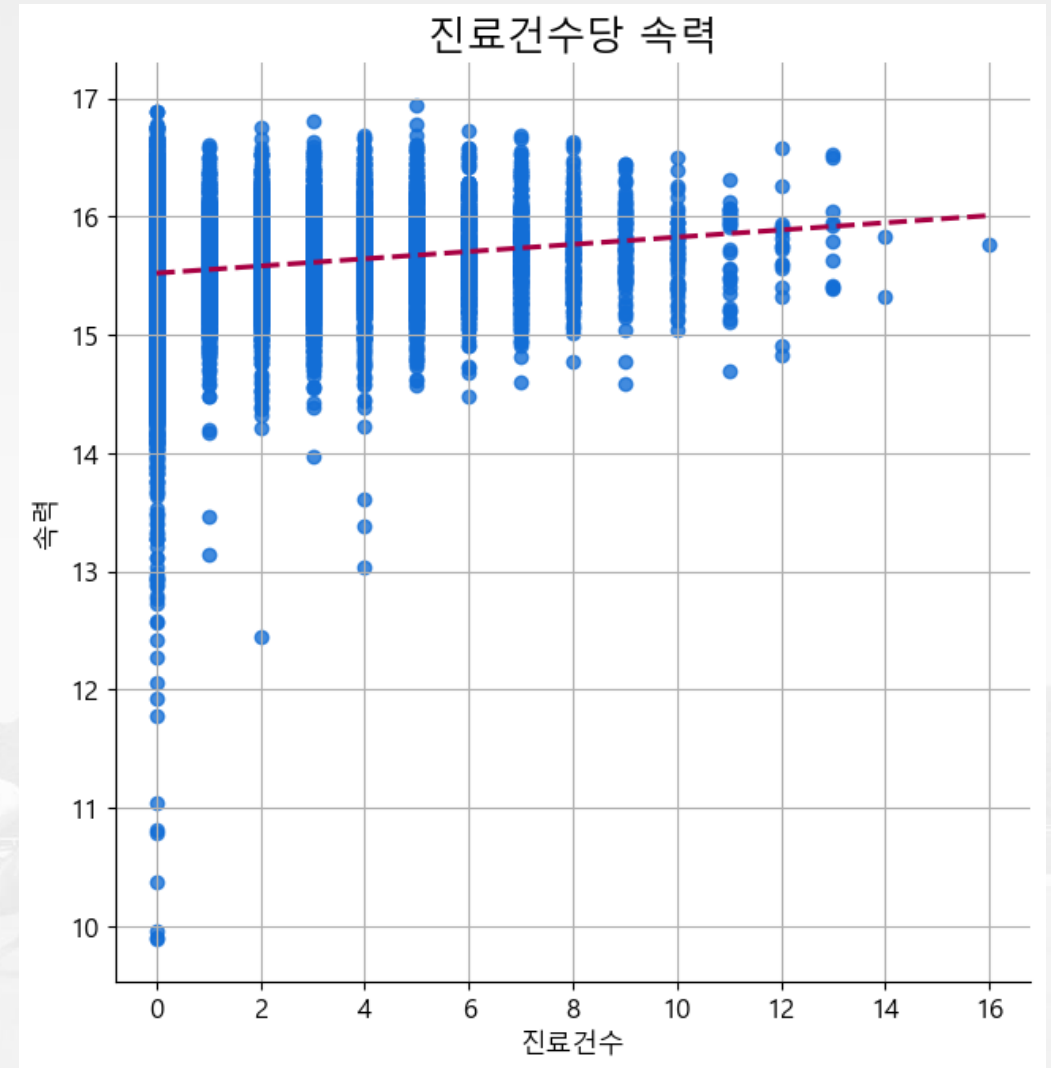
분석 결과 도출

- 데이터 필터링의 중요성
 - rank(경기 등급)를 '국6등급'으로 제한하지 않고 scatterplot과 추세선을 확인할 경우 부담 중량이 높아질수록 속도도 낮아지는 경향을 확인할 수 있음
 - 같은 티어의 경기가 아니므로 잘못된 분석 결과임
 - 올바른 분석 결과를 위해서는 데이터 필터링을 적절하게 해야 할 필요가 있음



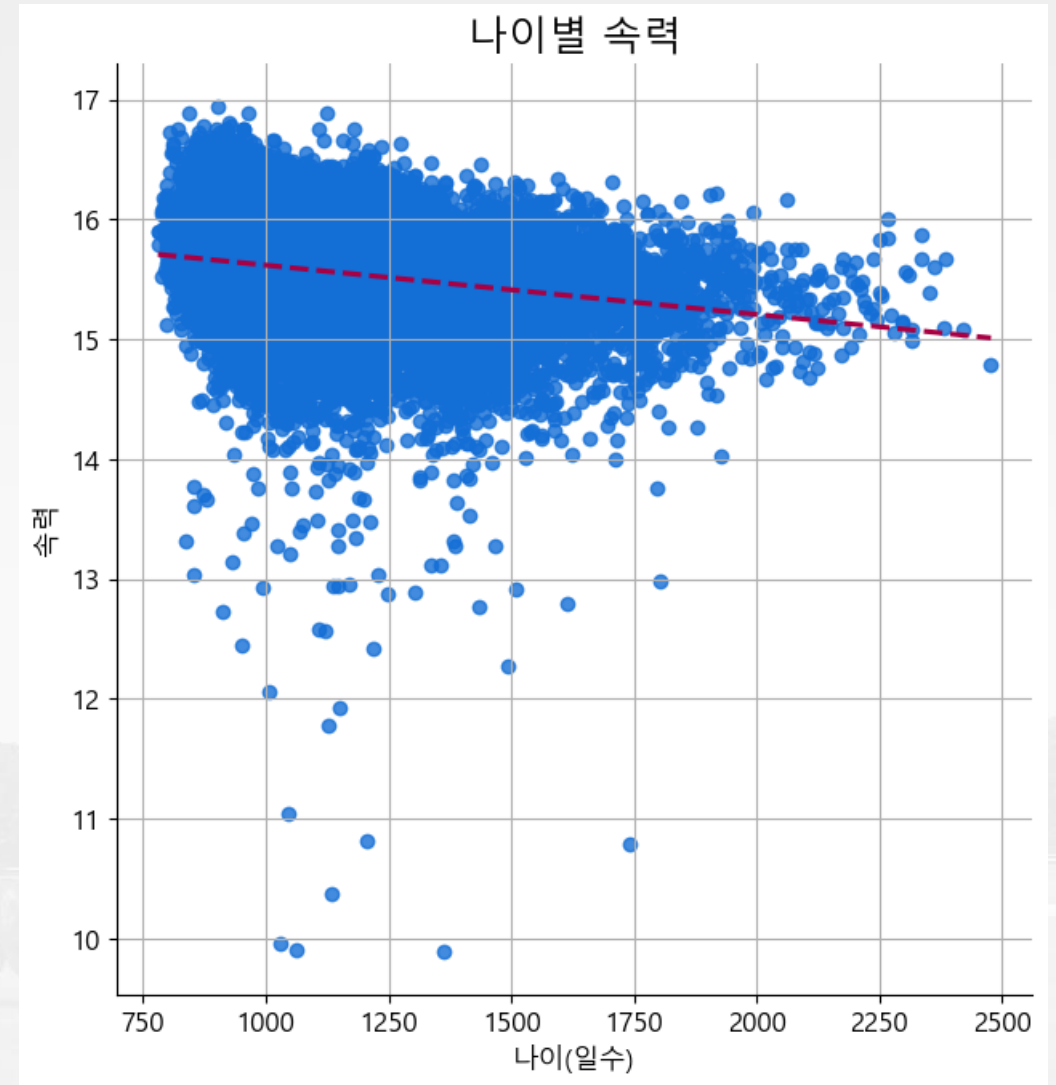
분석 결과 도출

- hrSpd와의 상관관계 시각화 – 진료건수
 - 피어슨 상관계수 = 0.14
 - Implot 사용
 - 진료건수가 증가할수록 말의 속력이 증가
- 일반적인 상식에 의한 결과가 아님
- 단순히 진료 건수로만 판단한 결과이므로 올바르지 않은 분석일 가능성이 높음
- 해결책 : 진료기록에 기록된 질병별로 가중치를 부여하여 시각화한다면 정확한 결과를 얻을 수 있을 것으로 추측됨



분석 결과 도출

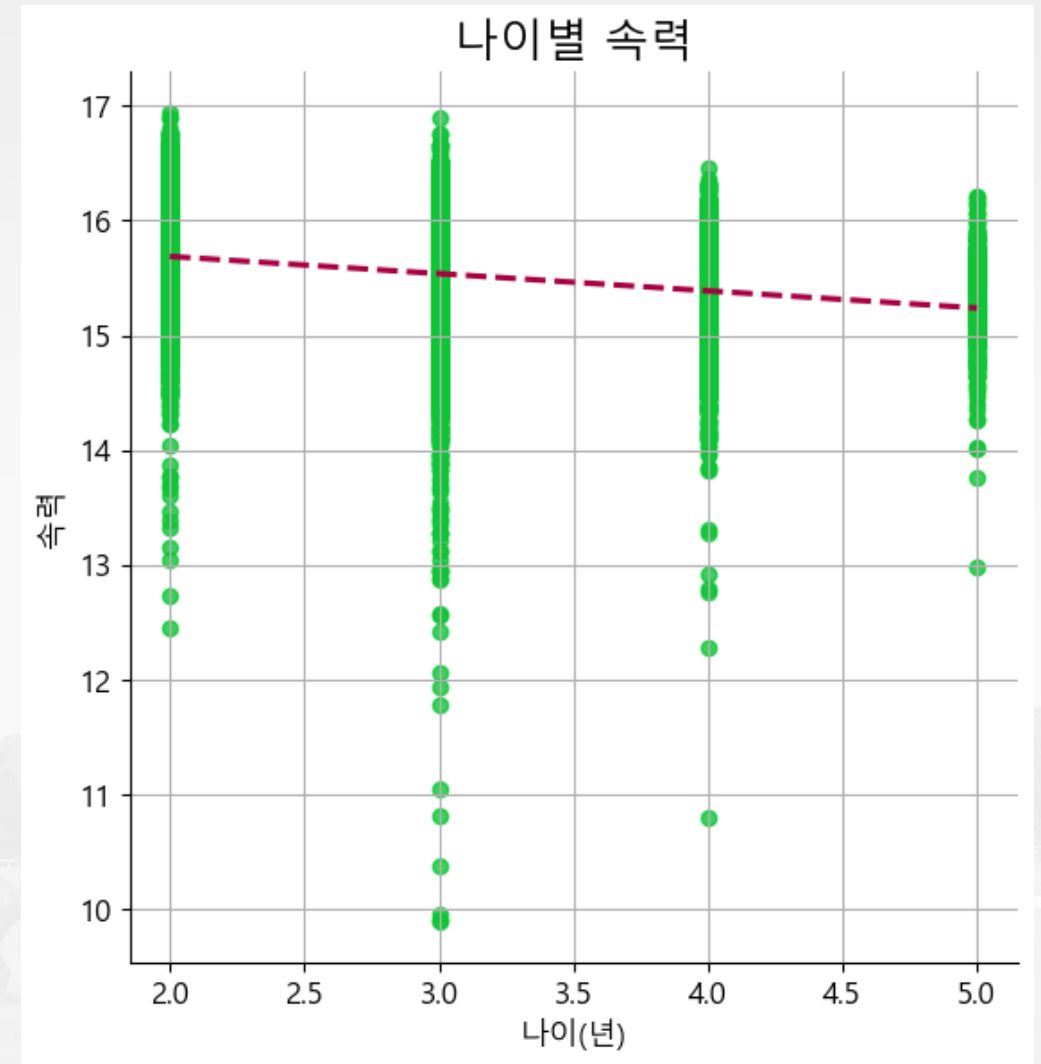
- hrSpd와의 상관관계 시각화 – 나이(일단위) (age_day)
 - 피어슨 상관계수 = -0.23
 - Implot 사용
 - 나이가 많을수록 속도도 감소하는 추세를 보임
- 상식에 의한 결과
- 노령화된 말은 경주 성적도 떨어지는 점을 확인할 수 있음



분석 결과 도출

- hrSpd와의 상관관계 시각화 – 나이(연단위) (age)
 - 피어슨 상관계수 = -0.26
 - Implot 사용
 - 나이가 많을수록 속력도 감소하는 추세를 보임

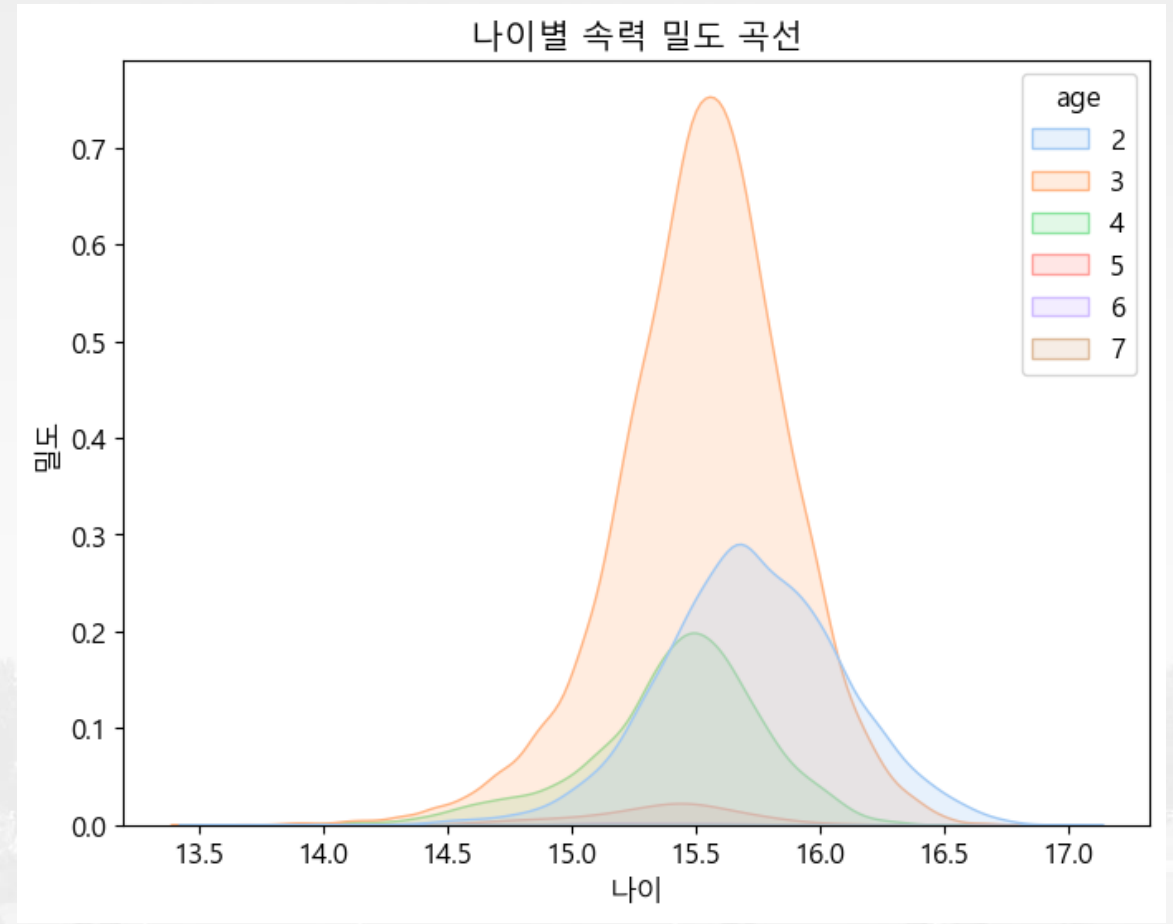
➤ 일단위 분석과 사실상 동일한 결과



분석 결과 도출

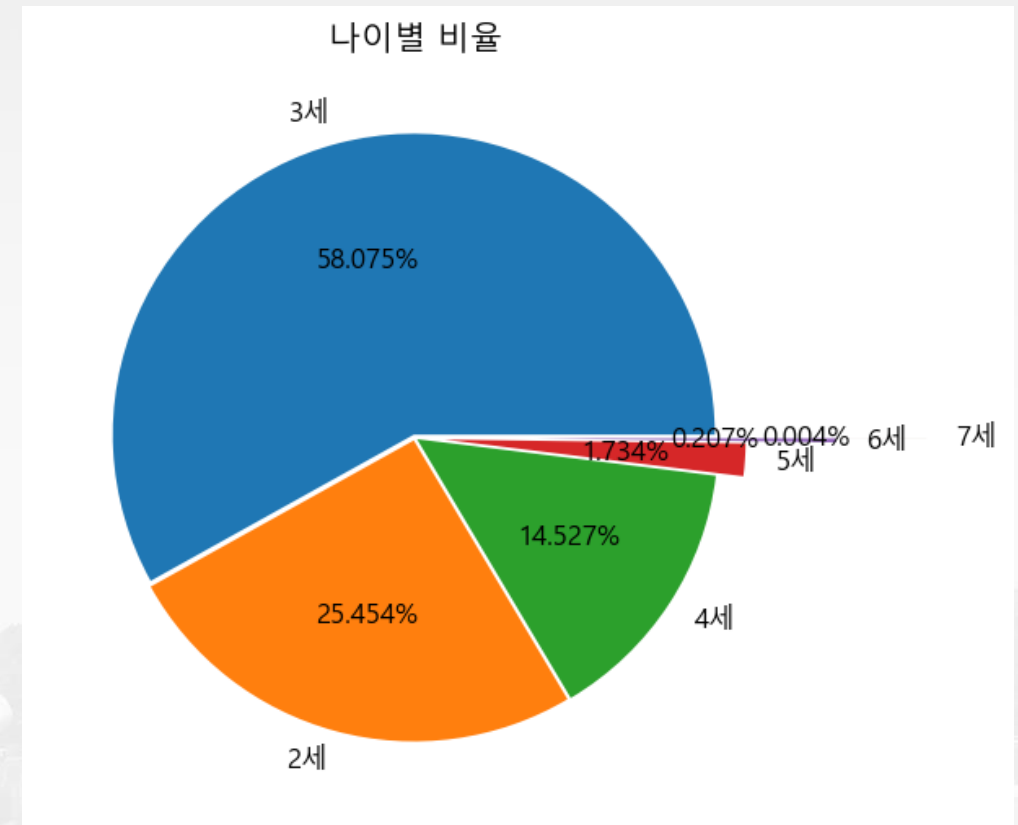
- hrSpd와의 상관관계 시각화 – 나이(연단위)
 - kdeplot 사용 (나이별로 분류)
 - 기술통계량 계산
- Lmplot과 마찬가지로 고령의 말일수록 속도 평균 및 중앙값이 감소하는 것을 확인할 수 있음

	count	mean	std	min	25%	50%	75%	max
2세	7135.0	15.715028	0.382242	12.448133	15.483871	15.719468	15.957447	16.949153
3세	16279.0	15.517455	0.378088	9.887006	15.325670	15.544041	15.748031	16.891892
4세	4072.0	15.411255	0.380430	10.791367	15.240328	15.457788	15.643803	16.460905
5세	486.0	15.338918	0.372860	12.987013	15.140764	15.399423	15.564202	16.216216
6세	58.0	15.324177	0.297262	14.675052	15.107052	15.311705	15.544041	16.009852
7세	1.0	14.783527	NaN	14.783527	14.783527	14.783527	14.783527	14.783527



분석 결과 도출

- 말의 나이 비율 확인
 - pieplot 사용
 - 3세가 절반 이상을 차지하며, 2 / 4세가 나머지의 대다수를 차지함
- 3세가 가장 메이저한 나이대, 마령이 5세를 넘어가면 기량 하락으로 인하여 대다수가 은퇴하는 것으로 추측됨



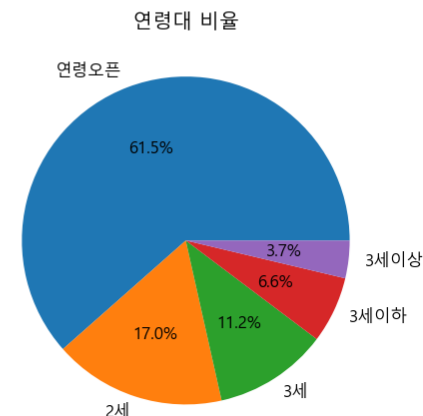
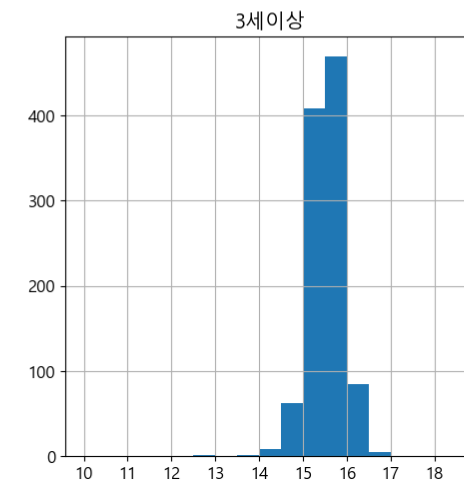
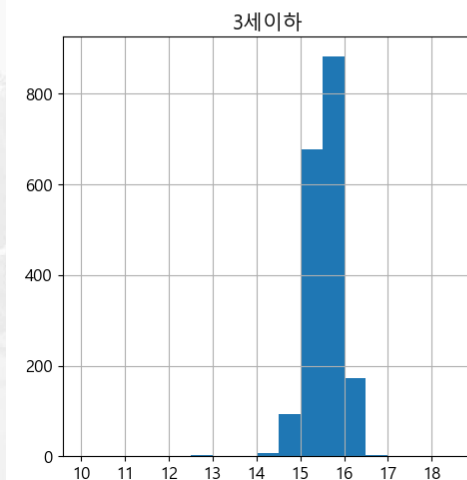
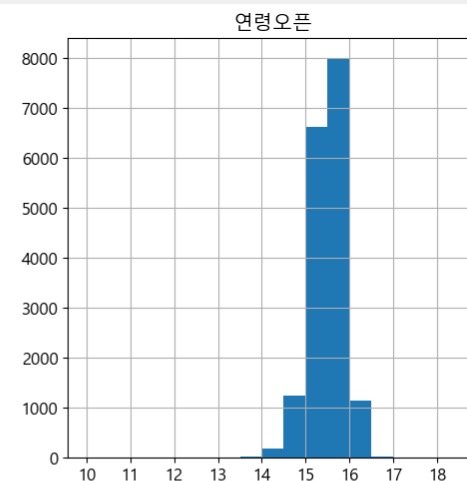
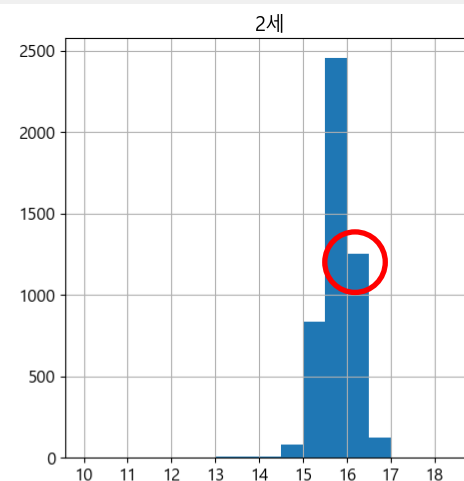
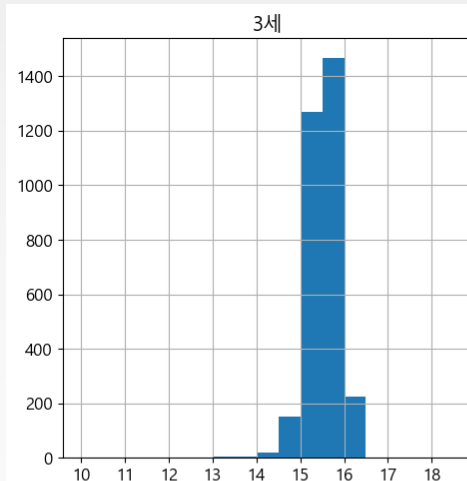
분석 결과 도출

• hrSpd와의 상관관계 시각화 – 나이제한 경기별(ageCond)

- histogram, pieplot 사용
- 기술통계량 계산

- 낮은 나이의 경기(특히 2세)에서 평균 속력이 훨씬 빠름
- 2세 속력 히스토그램상 16~16.5 속력이 비율이 타 나이대에 비해 확연히 높음

	count	mean	std	min	25%	50%	75%	max
3세	3145.0	15.511639	0.373271	10.371651	15.312132	15.527950	15.727392	16.528926
2세	4765.0	15.784070	0.379606	12.448133	15.552100	15.795869	16.025641	16.949153
연령오픈	17243.0	15.493146	0.380516	9.887006	15.306122	15.523933	15.727392	16.750419
3세이하	1839.0	15.541995	0.373973	10.810811	15.348288	15.564202	15.768725	16.891892
3세이상	1039.0	15.517857	0.374184	12.961117	15.306122	15.533981	15.748031	16.750419

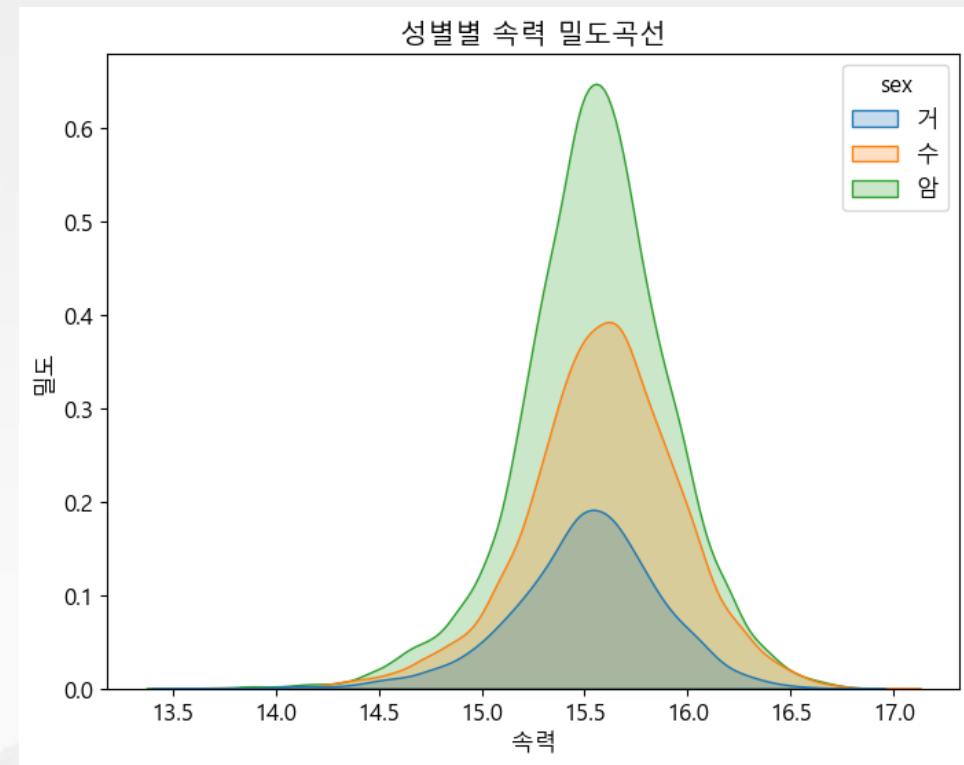


분석 결과 도출

- hrSpd와의 상관관계 시각화 – 성별(sex)

- kdeplot 사용
- 기술통계량 계산

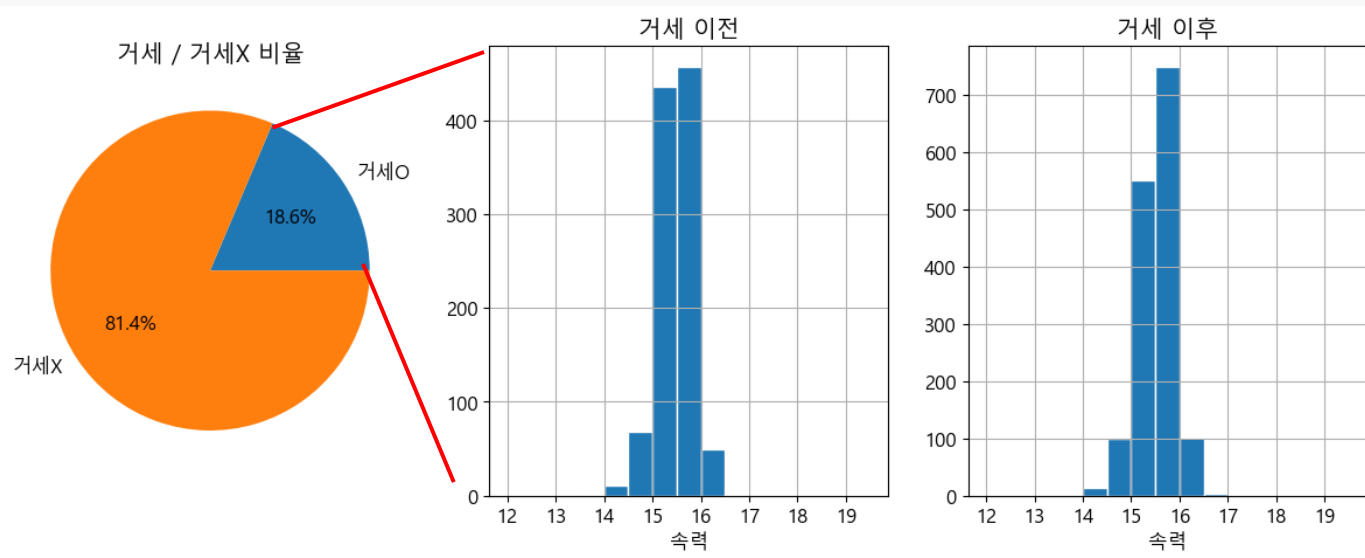
➤ 숫말 > 암말 > 거세마 순서로 평균속력이 우수함



	count	mean	std	min	25%	50%	75%	max
거세마	4273.0	15.498656	0.394807	9.887006	15.300546	15.523933	15.738499	16.750419
숫말	9416.0	15.579547	0.395042	10.810811	15.366430	15.594542	15.825915	16.949153
암말	14342.0	15.543546	0.390853	9.909166	15.345269	15.555556	15.772871	16.891892

분석 결과 도출

- 슷말 성적이 더 좋은데 거세를 하는 이유?
 - 모든 슷말이 거세를 하지는 않음
 - 거세를 한 슷말과 거세를 하지 않은 슷말을 나눔
 - 거세를 한 슷말 대상으로 거세 전후의 경기 기록을 비교
 - Pieplot, histogram 사용, 기술통계량 계산

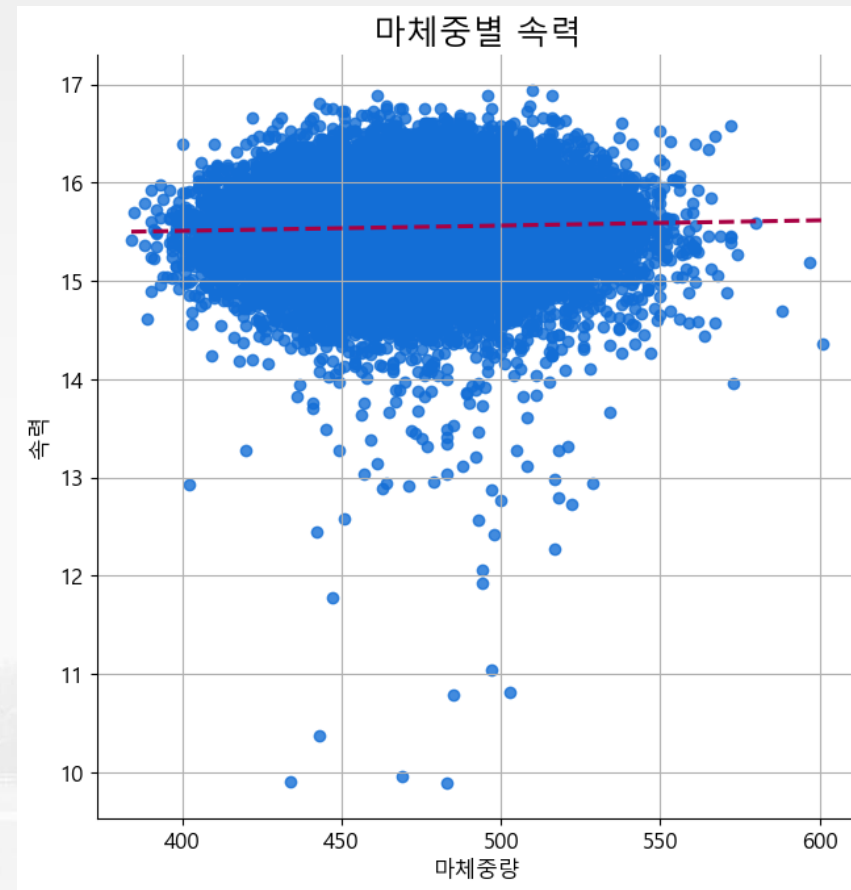
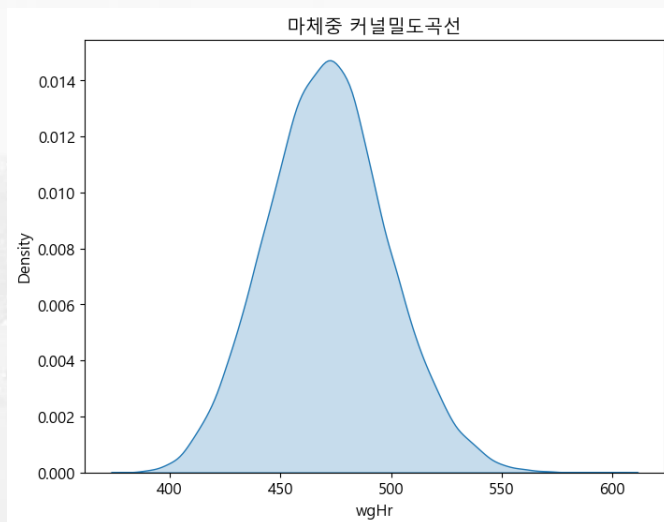


- 전체 슷말 중 18.6%가 거세를 함
- 거세한 말이 거세하기 전보다 성적이 상승
- 거세여부와 성적이 관련이 있다면 거세한 말이 성적이 좋지 않은 게 아니라 성적이 좋지 않아서 거세를 한 것

	count	mean	std	min	25%	50%
거세 전	1027.0	15.469031	0.382010	11.039558	15.282800	15.494636
거세 후	1519.0	15.510128	0.363171	12.422360	15.330189	15.544041

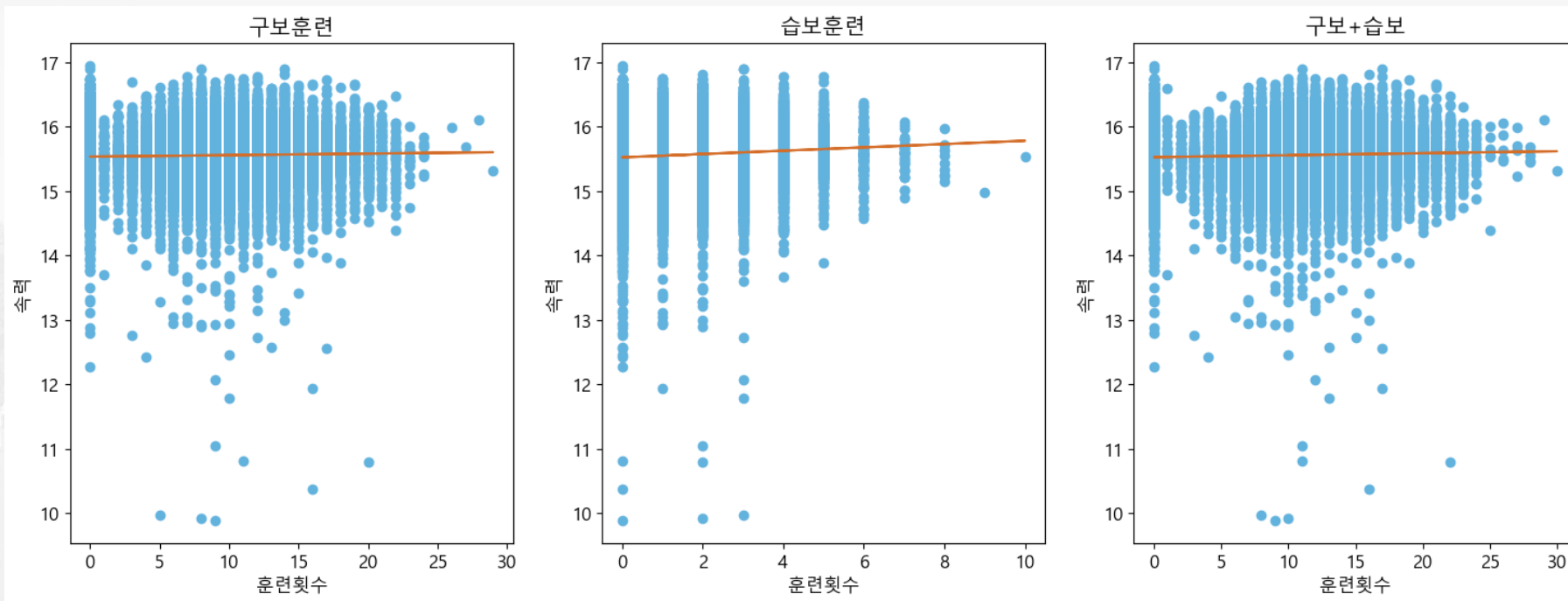
분석 결과 도출

- hrSpd와의 상관관계 시각화 – 마체중(wgHr)
 - 피어슨 상관계수 = 0.04
 - Implot 사용
 - 마체중과 속도 간에는 상관관계가 약함
- Scatter 모양이 타원형으로 나타나는 것은 400후 반대의 마체중의 분포가 가장 수가 많기 때문인 것으로 확인



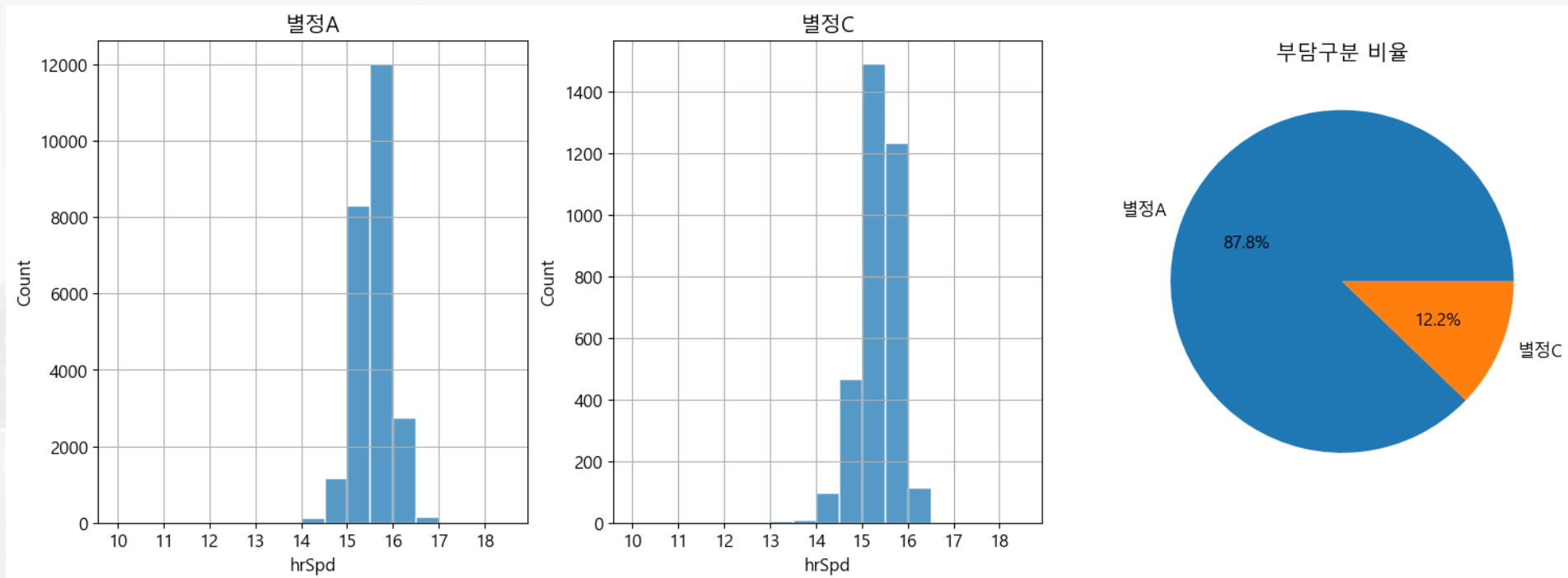
분석 결과 도출

- hrSpd와의 상관관계 시각화 – 훈련횟수(tr1, tr2)
 - 피어슨 상관계수 = tr1 : 0.03, tr2 = 0.08
 - Implot 사용
- 구보훈련은 성적과 연관성이 매우 약함
- 습보훈련은 성적 향상에 미미하지만 효과가 있음



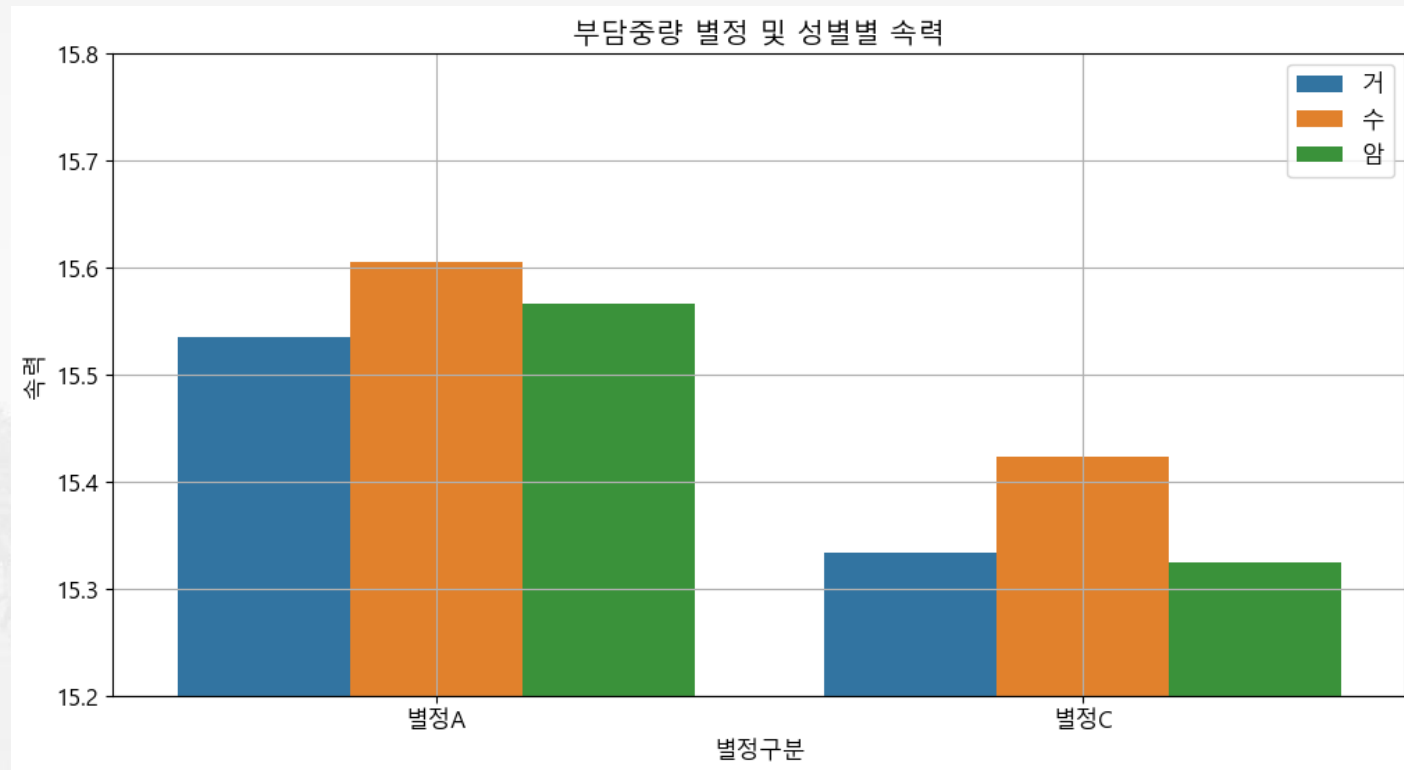
분석 결과 도출

- hrSpd와의 상관관계 시각화 – 부가중량 규정별(budam)
 - Histogram, pieplot 사용
 - 별정A가 87.8%, 별정C가 12.2% 비율
 - 별정A는 성/나이별 고정부담중량, C는 성별부담중량 추가보정 (암말은 부담중량 줄임)



분석 결과 도출

- hrSpd와의 상관관계 시각화 – 부가중량 규정별(budam)
 - Barplot 사용 (별정구분 및 성별별로 구분)
 - 별정A보다 별정C가 전반적인 성적이 떨어지며, 암말의 경우 별정C에서 거세마보다도 성적이 낮음
 - 암말에 부담중량을 더 적게 주는 규정인데 암말의 성적이 최하위권



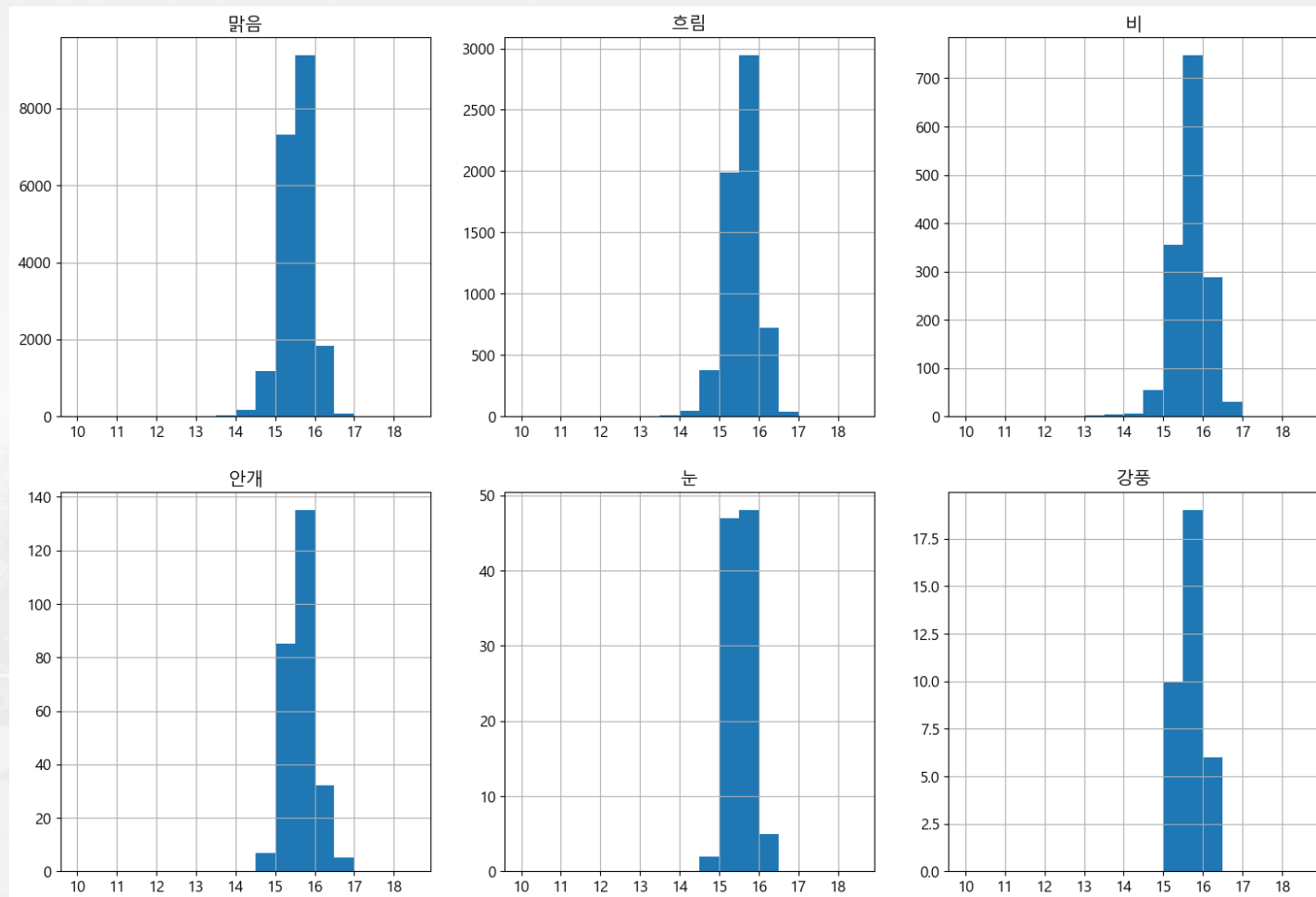
분석 결과 도출

- hrSpd와의 상관관계 시각화 – 날씨(weather)

- Histogram 사용, 기술통계량 계산

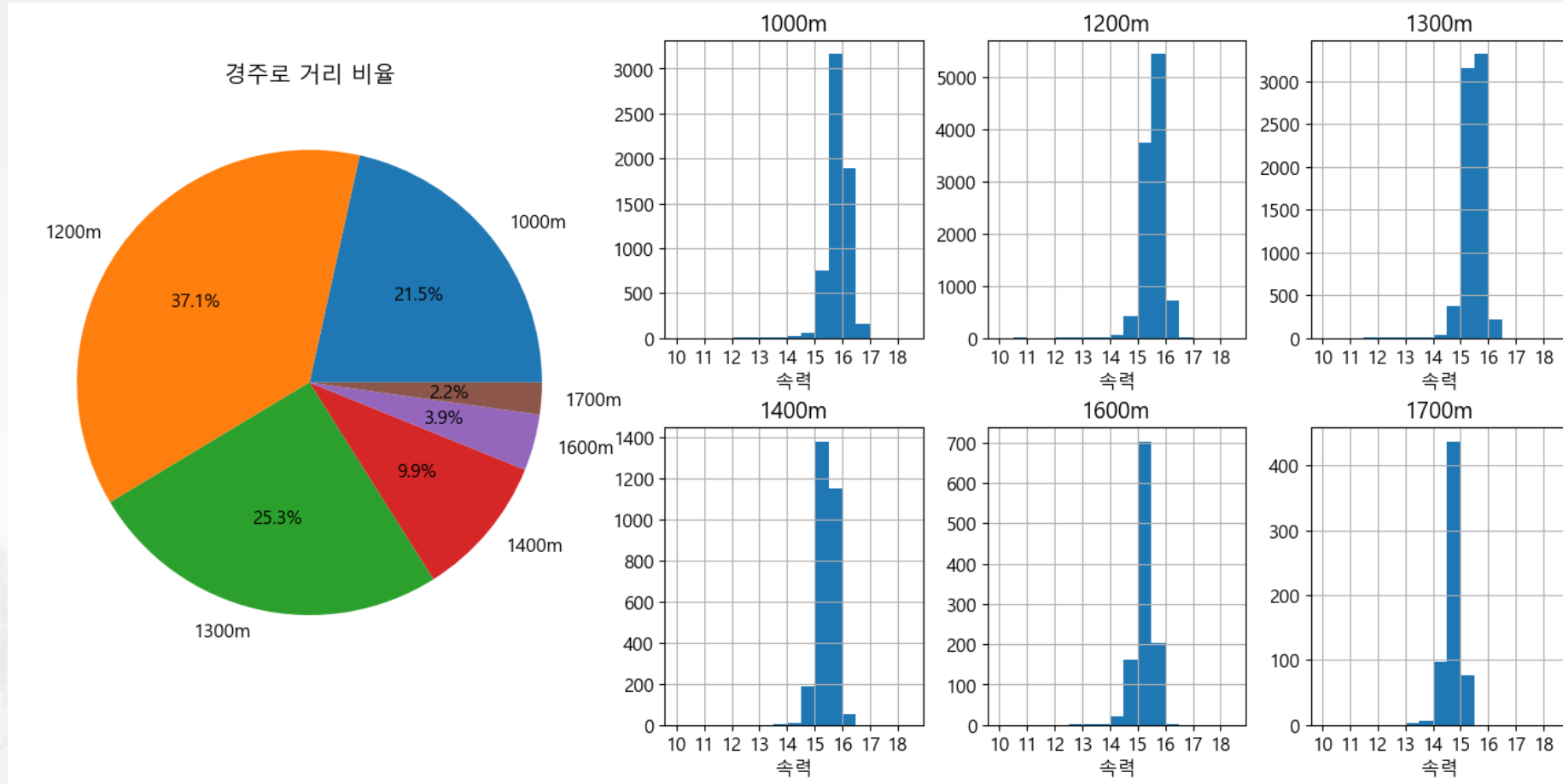
- 경주성적은 날씨와 큰 연관성이 없는 것으로 추정됨
- 오히려 우천시 경주성적이 평균으로는 가장 우수했음

	count	mean	std	min	25%	50%
맑음	20007.0	15.531421	0.385946	9.909166	15.325670	15.550239
흐림	6132.0	15.566037	0.409310	9.887006	15.364917	15.584416
비	1491.0	15.693603	0.412179	13.143483	15.463918	15.717092
안개	264.0	15.647961	0.348572	14.563107	15.415957	15.625000
눈	102.0	15.508935	0.277008	14.705882	15.307625	15.512470
강풍	35.0	15.659607	0.331588	15.081206	15.439430	15.673981



분석 결과 도출

- hrSpd와의 상관관계 시각화 – 경주로 거리(rcDist) : Histogram, Pieplot 사용



- 1000m ~ 1300m의 단거리 경주가 가장 많았음
- 1600m 이상의 장거리 경주는 속력이 점점 떨어지는 추세를 보임

분석 결과 도출

- hrSpd와의 상관관계 시각화 – 경주로 거리(rcDist)
 - 기술통계량 계산

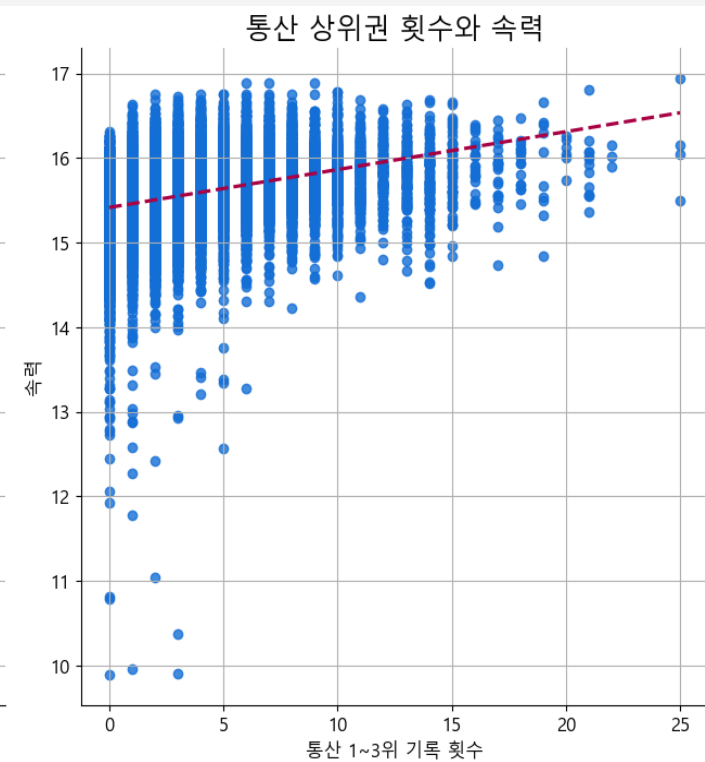
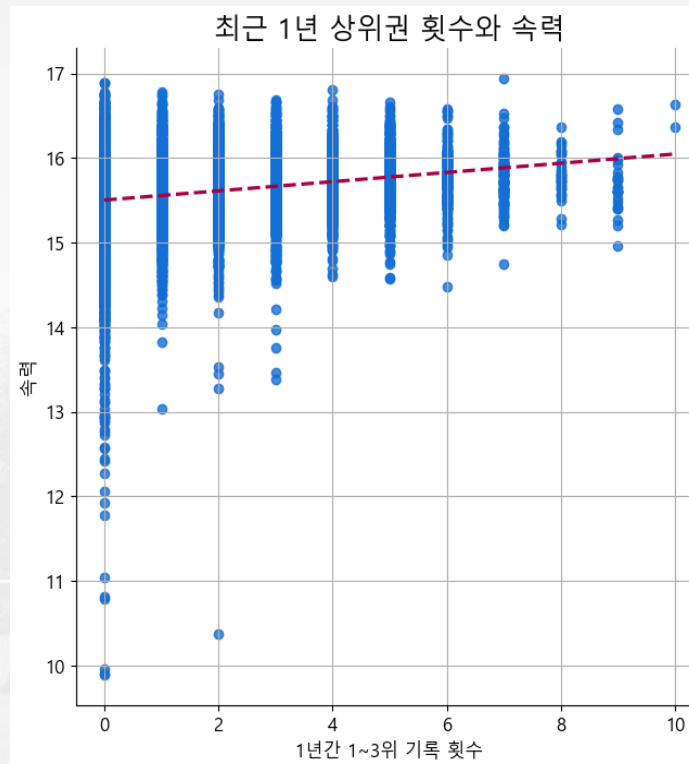
	count	mean	std	min	25%	50%	75%	max
1000m	6035.0	15.842644	0.357090	12.062726	15.625000	15.847861	16.077170	16.949153
1200m	10396.0	15.538105	0.356784	9.909166	15.345269	15.564202	15.748031	16.597510
1300m	7103.0	15.483449	0.311052	11.775362	15.312132	15.494636	15.681544	16.434893
1400m	2780.0	15.427616	0.309490	9.887006	15.267176	15.452539	15.625000	16.298021
1600m	1098.0	15.240797	0.326193	12.789768	15.094340	15.267176	15.444015	16.032064
1700m	619.0	14.703874	0.266321	13.270882	14.567266	14.718615	14.873141	15.412511

- 기술통계량으로 파악한 결과 경주거리가 짧을수록 말의 속력은 빨라지는 경향을 보임
- 사람의 육상경기와 유사한 추세

분석 결과 도출

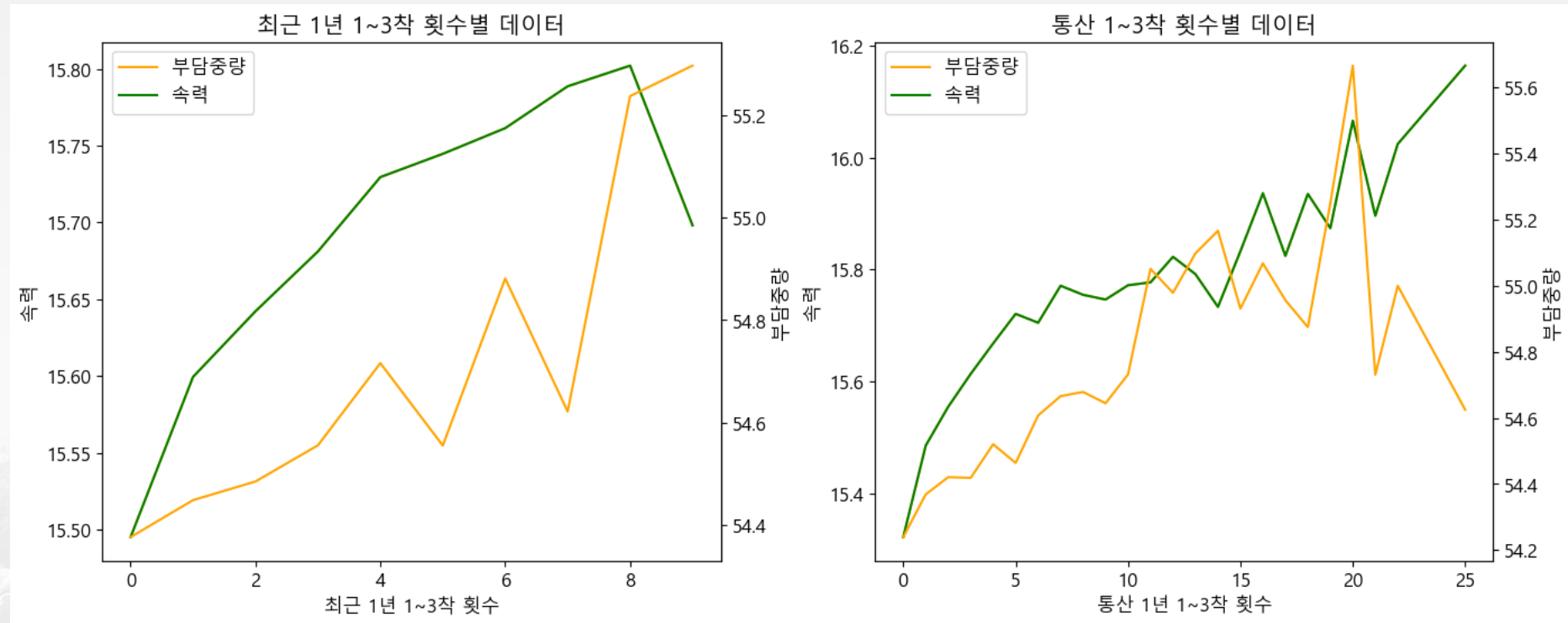
- hrSpd와의 상관관계 시각화 – 상위권 기록 횟수(topCntY, topCntT)
 - 피어슨 상관계수 = 최근 1년 기준 : 0.2, 통산 기준 : 0.35
 - Implot 사용

- 최근 1년, 통산 상위권 성적 모두 횟수가 많을수록 성적이 오르는 추세를 보임
- 우수한 성적에 따라서 부담 중량이 커질 수 있겠으나 잘 달리던 말이 꾸준히 잘 달리는 경향이 있음



분석 결과 도출

- hrSpd와의 상관관계 시각화 – 상위권 기록 횟수(topCntY, topCntT)
 - 상위권 기록 횟수별 속력 및 부담중량 lineplot 확인



- 성적이 좋을수록 속력과 부담중량 모두 늘어나는 경향을 보임 (마지막 데이터가 경향을 벗어나는 현상은 데이터 수에 의한 이슈로 파악됨)

가설 검증

- 잘 달리는 말의 요소를 추측하는 가설 검증

구분	가설	검증결과	설명
나이	늙은 말일 수록 경주 성적이 떨어진다	O	나이가 많을수록 평균 속력이 감소하는 추세가 명확하다
성별	숫말이 암말보다 경주 성적이 우수하다	O	숫말이 평균 속력에서 암말보다 뛰어나며 이는 성별 별 부담중량 보정이 가해져도 동일한 추세를 보임
체중	체중은 근육량과 관련이 있으므로 적정 수준 내에서 높은 체중을 보유한 말이 성적이 우수할 것이다	X	체중과 말의 속력은 뚜렷한 상관관계를 보이지 않았음

- 경기에 영향을 주는 외부 요소에 관한 가설 검증

구분	가설	검증결과	설명
날씨	악천후일 경우 전반적인 경기 성적이 하락한다	X	날씨와 관련없이 말의 속력은 일정한 편으로 확인됨
코스	코스 길이가 길 수록 전반적인 주행 능력이 떨어진다	O	코스 길이가 길수록 평균 속력이 감소함

문헌 검증

- 논문 1 : 모마의 통산 상금이 자식 말의 성적에 관여한다

부모마 수득 상금과 순위와의 상관관계가 높진 않지만 모마 수득 상금에 따른 순위는 랜덤 포레스트 결과 0.22의 상관관계가 도출된 것으로 보아 부마에 비해 모마와 경주마의 성적이 더 관계가 있는 것으로 보인다.
(허태성, 송민섭, 고동수, 2022)

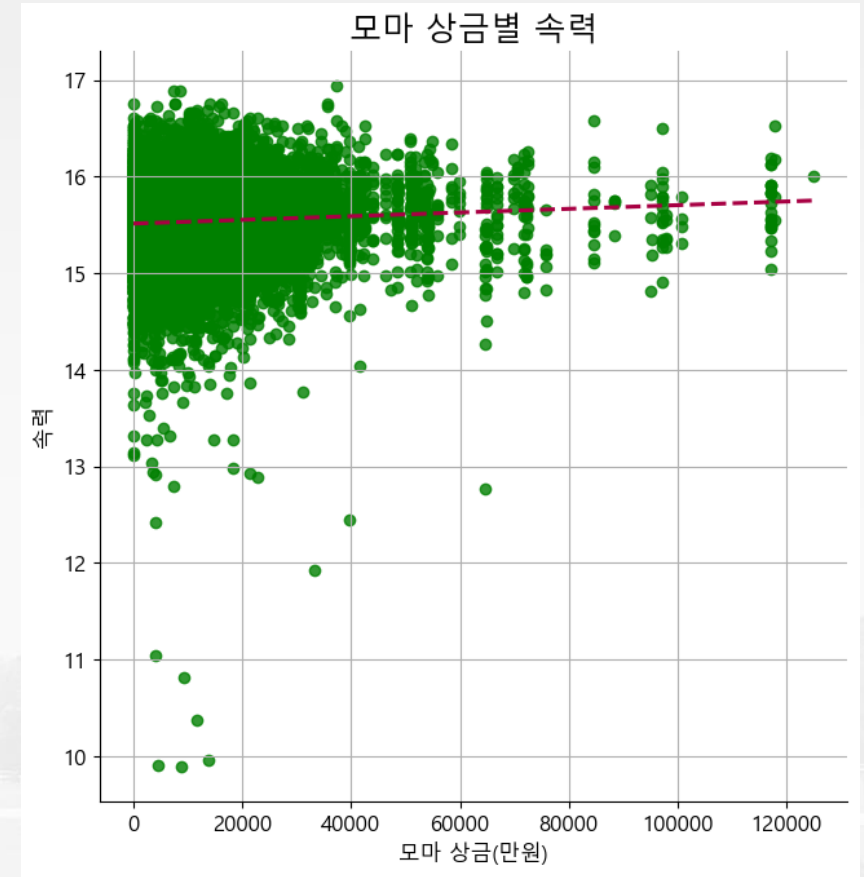
- 위 주장의 검증을 위하여 경주마 데이터에서 모마의 통산 상금을 계산한 컬럼을 생성 (moPrize, 단위는 만원)
 - 다만, 해외 말이 모마인 경우에는 데이터를 찾기 어려워 국내 모마로만 한정하여 데이터를 산출했으므로 데이터 손실이 발생



문헌 검증

- hrSpd와의 상관관계 시각화 – 모마의 상금(moPrize)
 - 피어슨 상관계수 = 0.06
 - Implot 사용
- 관련이 없진 않지만 상관관계는 약한 것으로 확인됨
- 논문에서는 종속변수를 순위(ord)로 지정하여 (Classification) 다른 결과가 도출된 것으로 추정됨

```
clf = RandomForestClassifier(n_estimators=100, max_depth=100, random_state=0)  
clf.fit(X,Y)
```



문헌 검증

- 논문 2 : 바깥 라인보다 안쪽 라인에서 출발하는 경주마의 성적이 더 좋다

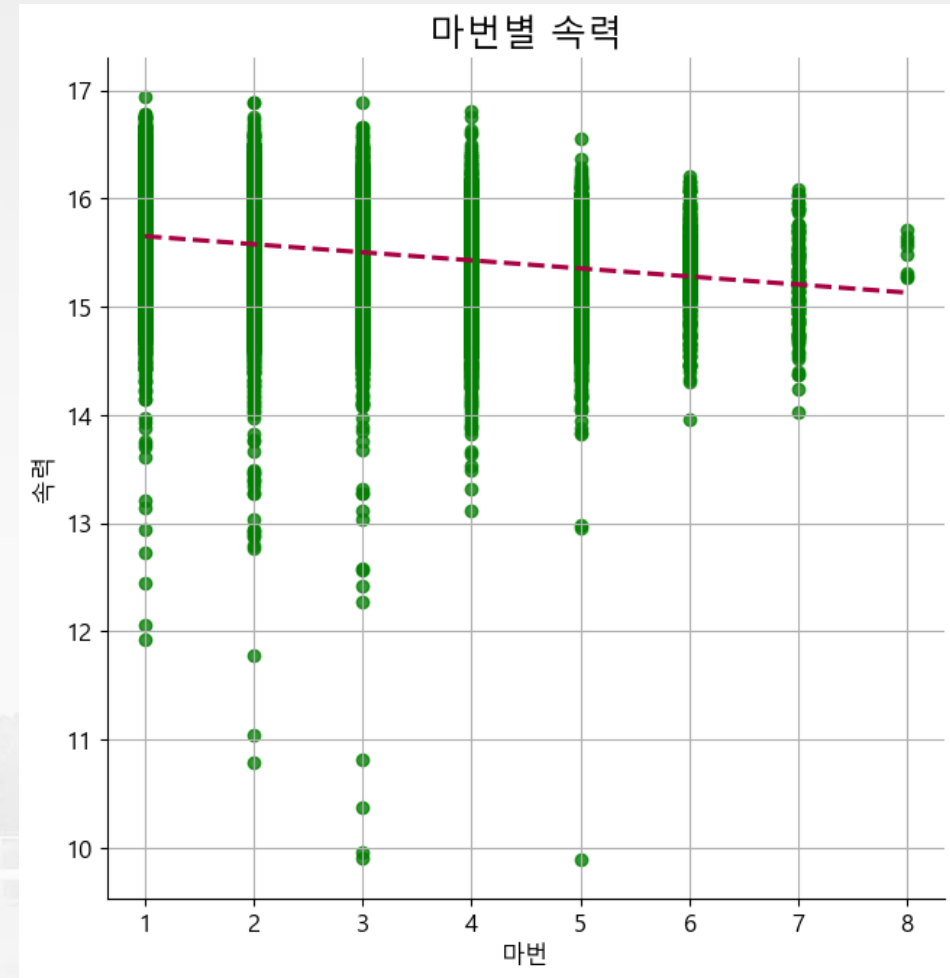
1000m ~ 1400m 경주의 경주전개는 직선-곡선-직선, 1700m ~ 2300m 경주의 경주전개는 직선-곡선-직선-곡선-직선으로 이루어져 있으므로 바깥 라인보다 안쪽 라인에서 출발하는 경주마의 기록이 더 좋을 것으로 예상된다.
(최혜민, 황나영, 황찬경, 송종우, 2022)

- 경기 데이터에 어느 라인에서 출발했는지에 관한 컬럼이 존재(rcNo)
 - rcNo를 마번으로 칭하며 숫자가 높을수록 바깥쪽에서 출발하게 됨(아웃코스)



문헌 검증

- hrSpd와의 상관관계 시각화 – 마번(rcNo)
 - 피어슨 상관계수 = -0.25
 - Implot 사용
- 음의 상관관계를 뚜렷하게 드러냄
- 마번이 낮은 인코스보다 마번이 높은 아웃코스보다 경주성적이 더 좋음
- 논문과 일치하는 결과

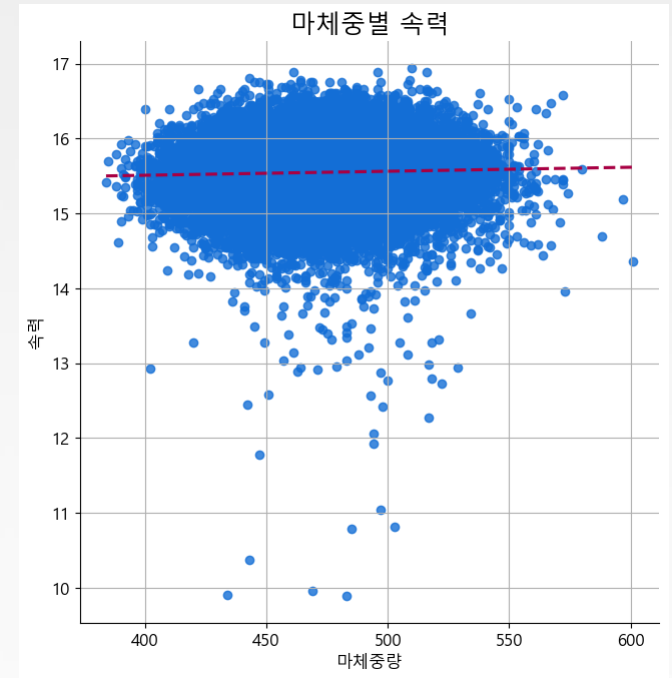


문헌 검증

- 논문 3 : 마체중이 경기 결과에 주는 영향이 크다 (김진홍, 2005)

경주거리	1000M	1200M	1400M	1700M	1800M	2000M
1차변수	경주기록	경주기록	경주기록	경주기록	경주기록	경주기록
2차변수	마체중	마체중	마체중	주로	마체중	부담중량

- 해당 주제에 대한 분석결과는 데이터 분석결과에 이미 포함되어 있음
 - 마체중은 속도 성적과 상관관계가 약하다
 - 해당 논문 또한 논문 1과 마찬가지로 사용한 머신러닝 모델이 로지스틱 회귀(사실상 분류모델) 및 의사결정나무 모델로 분류 분석이기 때문에 본 해석내용과 차이가 있을 수 있는 것으로 추정



결론

- 경마 성적에 영향을 주는 요소
 - 말 개별로 영향을 주는 요소와 경기 전체에 영향을 주는 요소로 구분하여 정리

구분	강한 상관관계	약한 상관관계 / 상관관계 없음
말 개별로 (순위예측)	나이 성별 상위권 기록 횟수 마번 (출발 위치)	부담중량 마체중 훈련 횟수 모마 상금
전역으로 영향 (경기 전체)	부가중량 규정 경주로 거리	날씨



결론

- 최종 결론

- 경마 성적에 영향을 주는 요소는 분명히 있었으며 완벽하지는 못하더라도 어느 정도의 예측이 가능한 스포츠로 볼 수 있다. (연령, 마번, 과거 성적 등)
- 부담 중량과 같은 경기 능력 보정 시스템 때문에 관람객의 예측을 흐리게 하는 요소가 있어 마권을 구매하는 관람객에 한정해 도박적인 요소 또한 동시에 띄고 있다.
- 관람객 입장
 - ✓ 우수한 말을 예측할 수 있는 연령, 과거 성적 통계 등의 요소를 고려하여 상위권을 어느정도 가늠을 잡을 여지가 있다
 - ✓ 다만 필승의 전략은 없으므로 베팅 손해를 최소화한다는 마인드로 접근이 필요
- 주최자 입장
 - ✓ 부담 중량 시스템을 좀 더 가다듬을 필요가 있다 (암말 보정이 암말에 불리하게 작용함)
 - ✓ 관람객이 예측하기 힘든 더 짜릿한 경기를 위해서는 경기 티어 등급을 세분화하는 것도 좋은 전략이 될 수 있음 (국6, 국5등급의 경기가 많으므로)

+ 04 과제 수행 후기

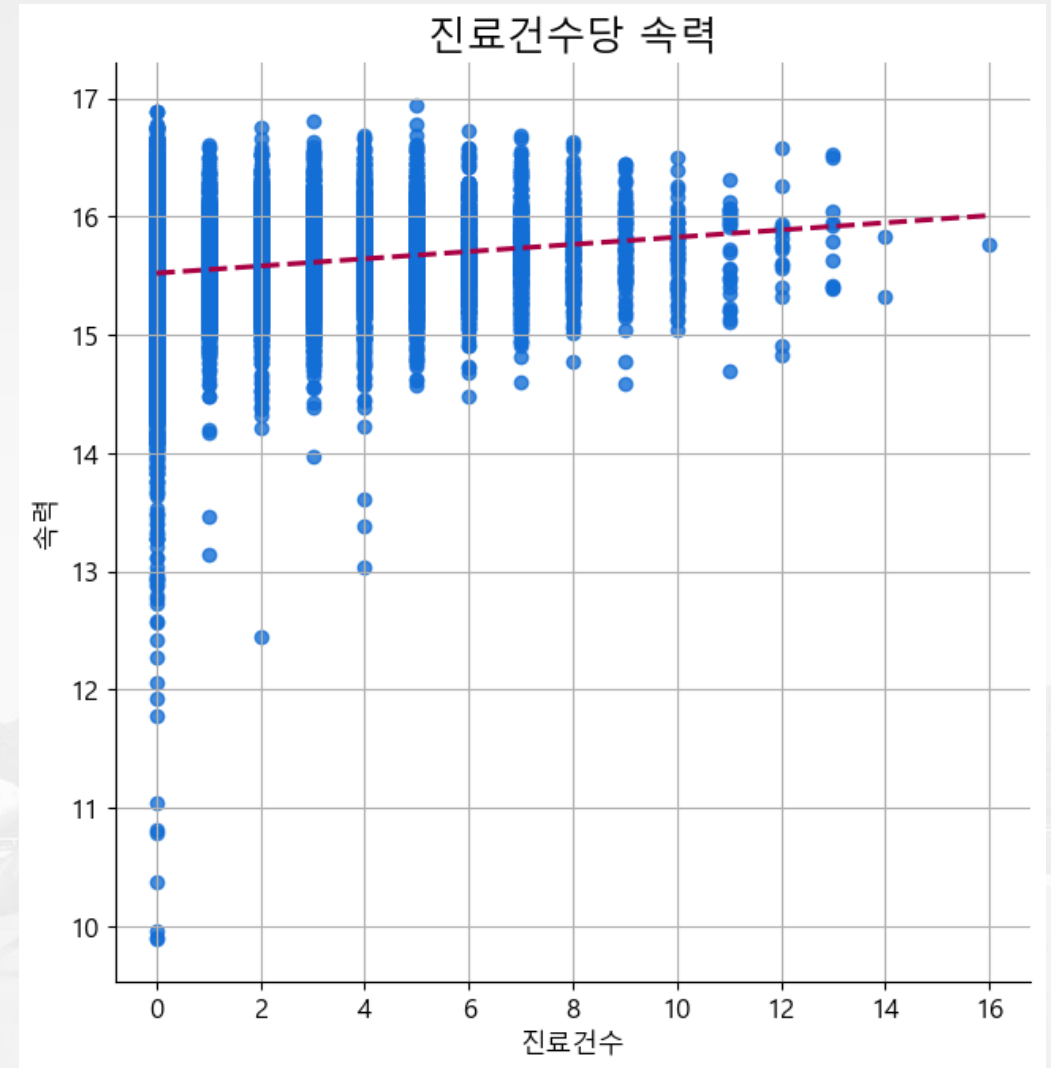
문제점 해결 과정 및 느낀 점

1. 문제점 해결 과정
2. 과제 진행 중 느낀 점



문제점 해결 과정

- 문제점 해결 과정 - 진료 데이터
 - 진료 데이터를 적합하게 사용할 방법을 찾지 못하였으며 최종적으로 진료건수를 이용하여 경주 성적과 상관관계를 찾으려 하였음
 - 결과 또한 좋지 않았음, 도메인 지식의 결여가 주요한 원인으로 확인됨
 - 단순 감기도 1건, 골절상도 1건으로 처리
 - 본문에서도 언급하였듯이 충분한 도메인 지식을 확보하여 질병 별로 가중치를 레이블 처리하여 결과 산출에 사용하면 좋은 결과를 얻어낼 수 있음



과제 진행 중 느낀 점

- 느낀 점

- 데이터 분석에 있어서 다각도의 시야가 매우 중요함을 느낌
 - 참고 문헌이나 자체적인 조사 과정에서 예상치 못한 요소가 종속변수에 영향을 줄 수 있다는 점을 확인하였음
- 수치 비교와 시각화 자료 비교를 적절하게 사용해야 함
 - 수치적으로는 파악할 수 없어 그래프를 이용해서 파악해야 했던 경우가 있었고 반대로 그래프로는 뚜렷한 데이터간 관계를 알 수 없어 기술통계량이나 상관관계수에 의존해야 하는 경우도 있었음
- 경마 자체에 관하여
 - 주관적으로는 경마는 도박에 가깝다고 여겼으나 그 역사가 깊고 말의 세부적인 사항에 따라 예측 가능한 요소가 있는 스포츠라는 사실을 알 수 있었음 (도박 요소가 없다는 것은 아님)
 - 관람객들이 지나치게 베팅에 몰두하지 않고 가볍게 즐긴다면 보다 건전한 스포츠로써의 인식이 확대될 수 있을 것으로 기대함

05 + 마무리

참고 문헌 및 Q&A

1. 참고 문헌

2. Q & A



참고 문헌

[논문]

- 47p : 허태성, 송민섭, and 고동수. “회귀 분석을 통한 경마 순위 예측 모형.” 한국컴퓨터정보학회 하계학술대회 논문집 30, no. 2 (2022): 2.
<https://scienceon.kisti.re.kr/commons/util/originalView.do?cn=CFKO202232249518440&oCn=NPAP13595101&dbt=CFKO&journal=NPRO00386978>.
- 49p : 최혜민, 황나영, 황찬경, and 송종우. “서울 경마 경기 우승마 예측 모형 연구.” 응용통계연구 28, no. 6 (2015): 4.
<https://scienceon.kisti.re.kr/commons/util/originalView.do?cn=JAKO201504255079117&dbt=JAKO&koi=KISTI1.1003%2FJNL.JAKO201504255079117>.
- 51p : 김진홍. “의사결정트리 프로세스를 이용한 경마순위 예측시스템에 대한 연구.” 서강대학교, 2005, 37.
<https://scienceon.kisti.re.kr/commons/util/originalView.do?cn=JAKO201504255079117&dbt=JAKO&koi=KISTI1.1003%2FJNL.JAKO201504255079117>.

[경마 관련 규정]

- 한국마사회, “서울경마_경마정보”, 한국마사회 경마정보, 2014, <https://race.kra.co.kr/seoulMain.do>.

[경마 데이터]

- 한국마사회, “한국마사회 경주기록 정보”, 공공데이터포털, 2023, <https://www.data.go.kr/data/15058305/openapi.do>.

Q & A

