경마에서 어떤 말이 잘 달릴까?

TEAM 3조 (🦉)

오용석
이정인
원석재

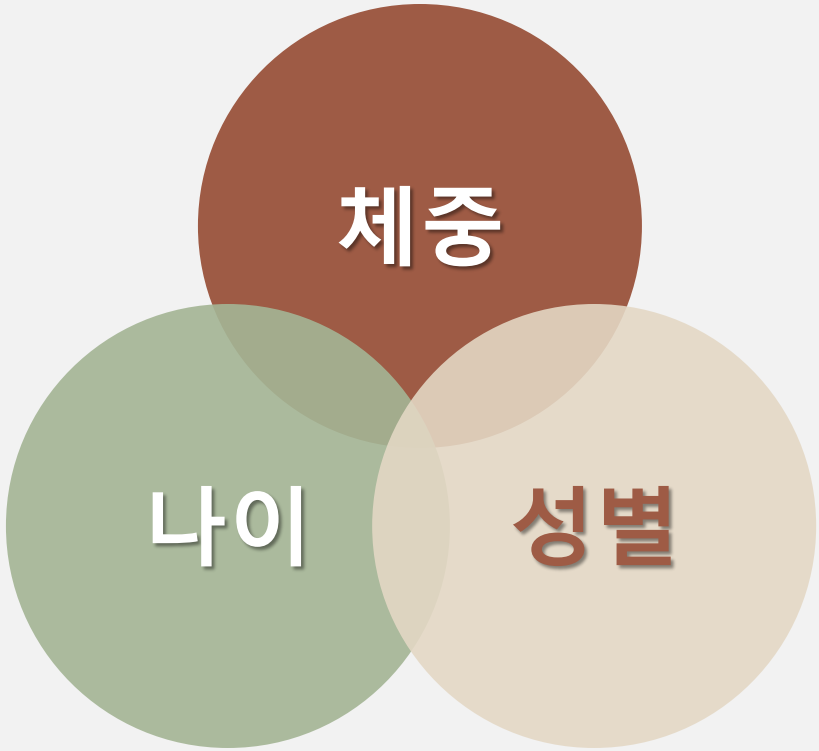
CONTENTS

01 프로젝트 개요

1등인 말을 맞추려면
어떤 특징을 가진 말을 골라야 할까?

경마를 하나도 모르는 세 명이 경마의 데이터만 가지고 어떤 말이 1등 할지 예측할 수 있는지에 대한 궁금증에서 시작한 프로젝트

경마의 다양한 요소 중에서 어떤 요소가 순위에 중요한 역할을 하는지 맞춰보자



02 프로젝트 팀 구성 및 역활

훈련생	역할	담당 업무
오용석	팀원	• 데이터가공 및 분석 • 시각화 및 연관관계 도출
이정인	팀원	• 날씨 별 1등을 가장 많이 한 말의 이름 • 거리 별 1등을 가장 많이 한 말의 이름
원석재	팀원	• 데이터 분석 & 활용

03 프로젝트 수행 절차 및 과정

■ 데이터 수집

기간:20190104 - 20230702

MEET 경마장	RC_DATE 경주일자	RC_NO 경주번호	RC_DIST 경주거리	RANK 경주등급	WEATHER 날씨	TRACK 주로상태(함수율)	ORD 착순	CHUL_NO 출주번호	HR_NAME 마명	HR_NO 마번	NAME 출신국가	PRD_NAME 국산외산구분	AGE 마필연령	SEX 마필성별	RC_TIME 경주기록(초)
제주	20230701	1	800	제6등급	흐림	포화 (18%)	1	9	한백제왕	3101848	한국	국산	3	거	65.2
제주	20230701	1	800	제6등급	흐림	포화 (18%)	2	2	명의로운	3102087	한국	국산	3	암	67.3
제주	20230701	1	800	제6등급	흐림	포화 (18%)	3	3	번개의꿈	3101489	한국	국산	4	수	67.5
제주	20230701	1	800	제6등급	흐림	포화 (18%)	4	6	고차원	3103528	한국	국산	2	암	67.6
제주	20230701	1	800	제6등급	흐림	포화 (18%)	5	1	위너고	3102911	한국	국산	3	암	67.7
제주	20230701	1	800	제6등급	흐림	포화 (18%)	6	5	백프로	3102675	한국	국산	3	수	67.8
제주	20230701	1	800	제6등급	흐림	포화 (18%)	7	4	성공가두	3101563	한국	국산	3	수	68.8
제주	20230701	1	800	제6등급	흐림	포화 (18%)	8	8	가화신화	3102088	한국	국산	3	암	68.8
제주	20230701	1	800	제6등급	흐림	포화 (18%)	9	7	여왕별	3102676	한국	국산	3	암	69.3
제주	20230701	1	800	제6등급	흐림	포화 (18%)	10	10	신으뜸	3102432	한국	국산	3	암	73.4
제주	20230701	2	800	제6등급	흐림	포화 (18%)	1	5	비룡신공	3102610	한국	국산	3	거	66.9
제주	20230701	2	800	제6등급	흐림	포화 (18%)	2	8	일급절주	3102690	한국	국산	3	거	66.9
제주	20230701	2	800	제6등급	흐림	포화 (18%)	3	1	무인수	3102598	한국	국산	3	암	67.4
제주	20230701	2	800	제6등급	흐림	포화 (18%)	4	3	십시일반	3102425	한국	국산	3	암	67.5
제주	20230701	2	800	제6등급	흐림	포화 (18%)	5	2	중문보스	3102843	한국	국산	4	거	67.7
제주	20230701	2	800	제6등급	흐림	포화 (18%)	6	6	대성여신	3101698	한국	국산	4	암	67.9
제주	20230701	2	800	제6등급	흐림	포화 (18%)	7	9	신조	3103719	한국	국산	2	수	68
제주	20230701	2	800	제6등급	흐림	포화 (18%)	8	10	호두	3102705	한국	국산	3	암	68.6
제주	20230701	2	800	제6등급	흐림	포화 (18%)	9	4	얼음새	3102604	한국	국산	3	거	69.4
제주	20230701	2	800	제6등급	흐림	포화 (18%)	10	7	캡틴전설	3101706	한국	국산	4	수	71
제주	20230701	3	900	제5등급	안개	포화 (18%)	1	3	바람아	3101736	한국	국산	3	암	74.5
제주	20230701	3	900	제5등급	안개	포화 (18%)	2	9	서귀보배	3102015	한국	국산	3	암	75
제주	20230701	3	900	제5등급	안개	포화 (18%)	3	7	누마루	3102119	한국	국산	4	암	75.3
제주	20230701	3	900	제5등급	안개	포화 (18%)	4	2	황해장군	3018715	한국	국산	5	거	75.6
제주	20230701	3	900	제5등급	안개	포화 (18%)	5	5	엄부탕진	3101804	한국	국산	3	암	75.7
제주	20230701	3	900	제5등급	안개	포화 (18%)	6	4	당산보스	3102272	한국	국산	3	거	77
제주	20230701	3	900	제5등급	안개	포화 (18%)	7	8	신전비상	3102956	한국	국산	3	암	77.1
제주	20230701	3	900	제5등급	안개	포화 (18%)	8	1	백야명성	3102569	한국	국산	4	암	77.2
제주	20230701	3	900	제5등급	안개	포화 (18%)	9	6	번영	3101712	한국	국산	4	암	78.3
제주	20230701	4	1000	제5등급	흐림	포화 (18%)	1	3	여신강림	3018870	한국	국산	5	암	84.7

(63293, 92)

03 프로젝트 수행 절차 및 과정

■ 데이터 전처리

○ 데이터 분리 (체중 및 트랙 상태)

- 체중 : 300 (-18) -> 300, -18
- 트랙 상태 : 포화(18%) -> 포화, 18

○ 중복 데이터, 실격 및 OPEN 경기 삭제

- ORD : 91 – 99 까지 실격 데이터
- OPEN : 제한이 없는 경기
- 중복 데이터 : 동일 데이터 반복 삽입 확인

```
df[['horse_weight', 'weight_gain', 'WG_HR', 'TR_CON', 'TR_WC']]
```

0.0s

	horse_weight	weight_gain	WG_HR	TR_CON	TR_WC
0	311.0	-18.0	311(-18)	포화	18
1	283.0	-6.0	283(-6)	포화	18
2	271.0	-5.0	271(-5)	포화	18
3	265.0	-8.0	265(-8)	포화	18
4	265.0	-2.0	265(-2)	포화	18
...	...	...	...	...	...
63288	504.0	6.0	504(+6)	건조	3
63289	494.0	-9.0	494(-9)	건조	3

Fig1. Data Splitting

```
drop_duplicate = selected_ord.groupby(['MEET', 'RC_DATE', 'RC_NO', 'ORD', 'HR_NAME'])['
drop_duplicate2 = drop_duplicate.reset_index()
len(drop_duplicate2[drop_duplicate2['RANK'] > 1].index)
```

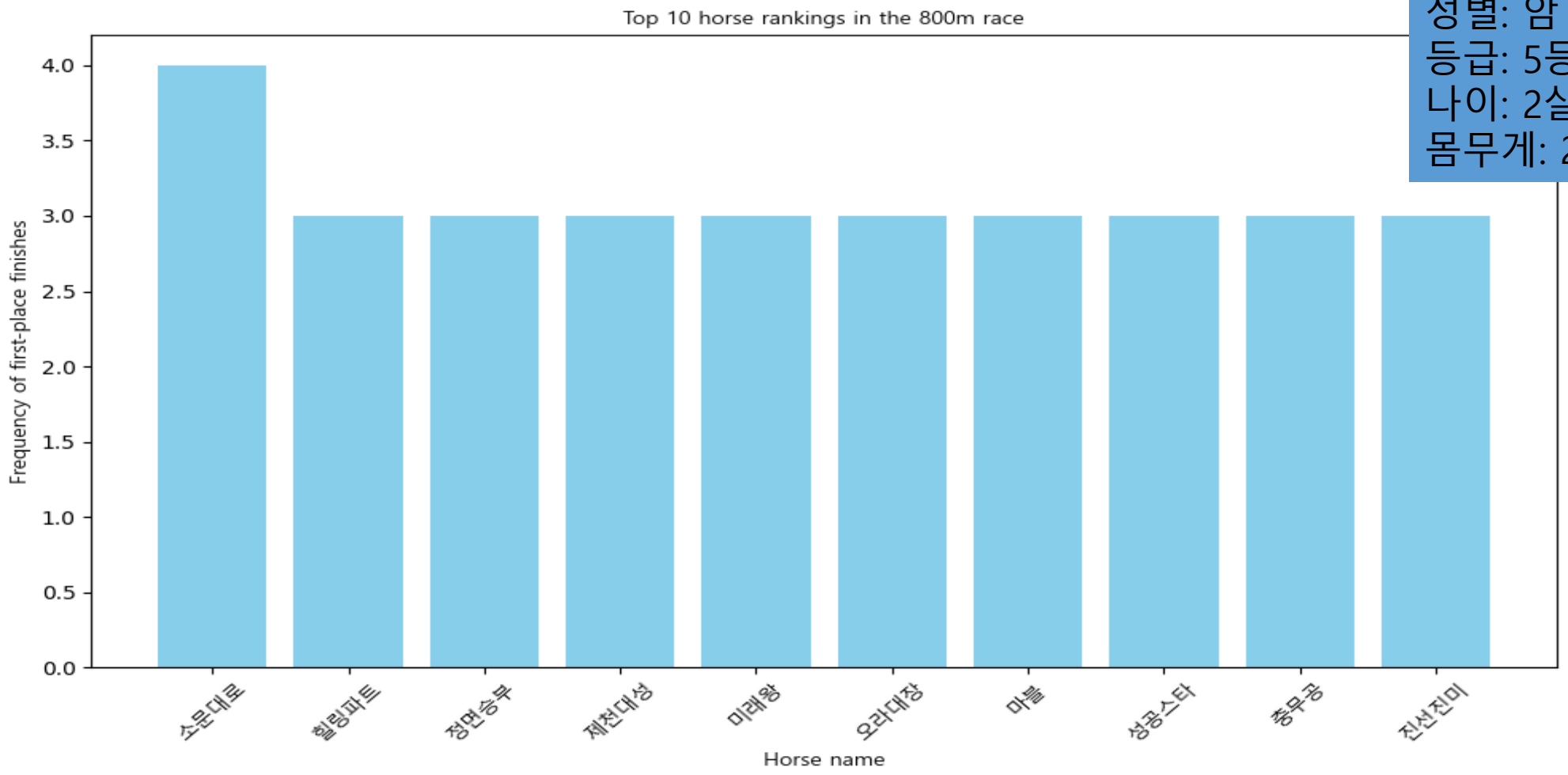
Python

```
0
```

Fig2. Removing Duplicates

03 프로젝트 수행 절차 및 과정

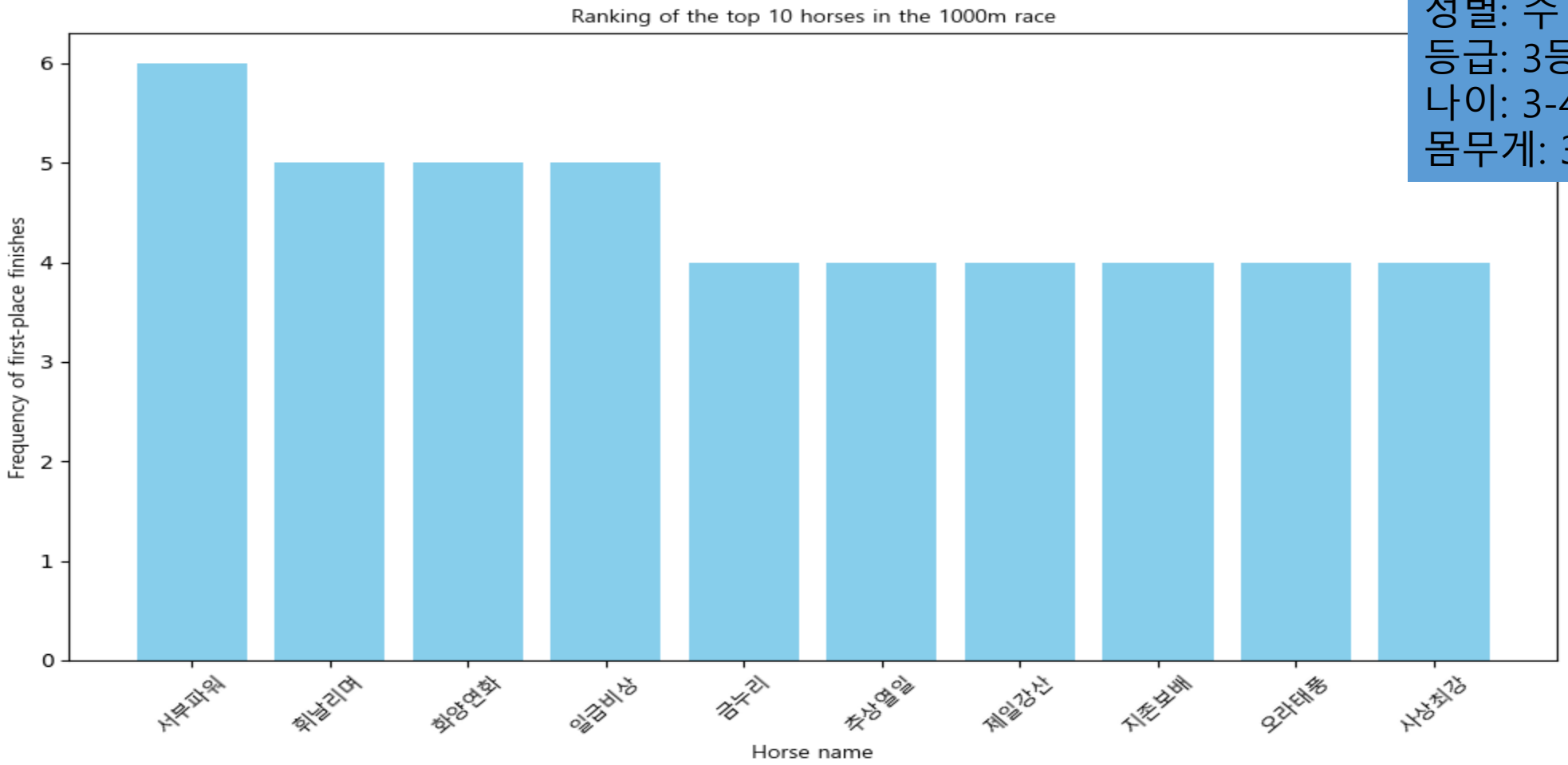
■ 거리 별 1등을 가장 많이 한 말의 이름 (800m)



이름: 소문대로
성별: 암
등급: 5등급
나이: 2살
몸무게: 285.5kg

03 프로젝트 수행 절차 및 과정

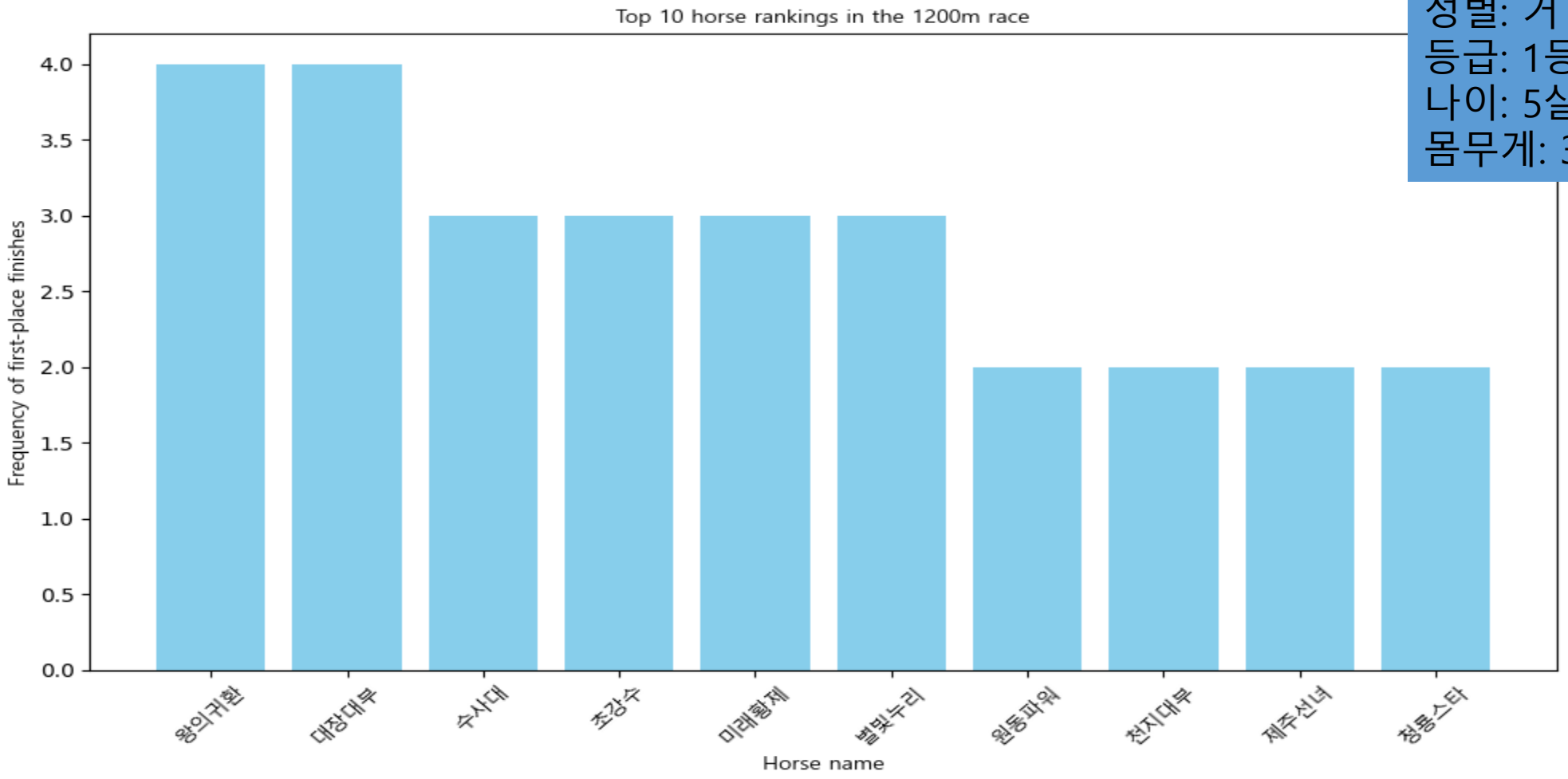
■ 거리 별 1등을 가장 많이 한 말의 이름 (1000m)



이름: 서부파워
성별: 수
등급: 3등급
나이: 3-4살
몸무게: 312.5kg

03 프로젝트 수행 절차 및 과정

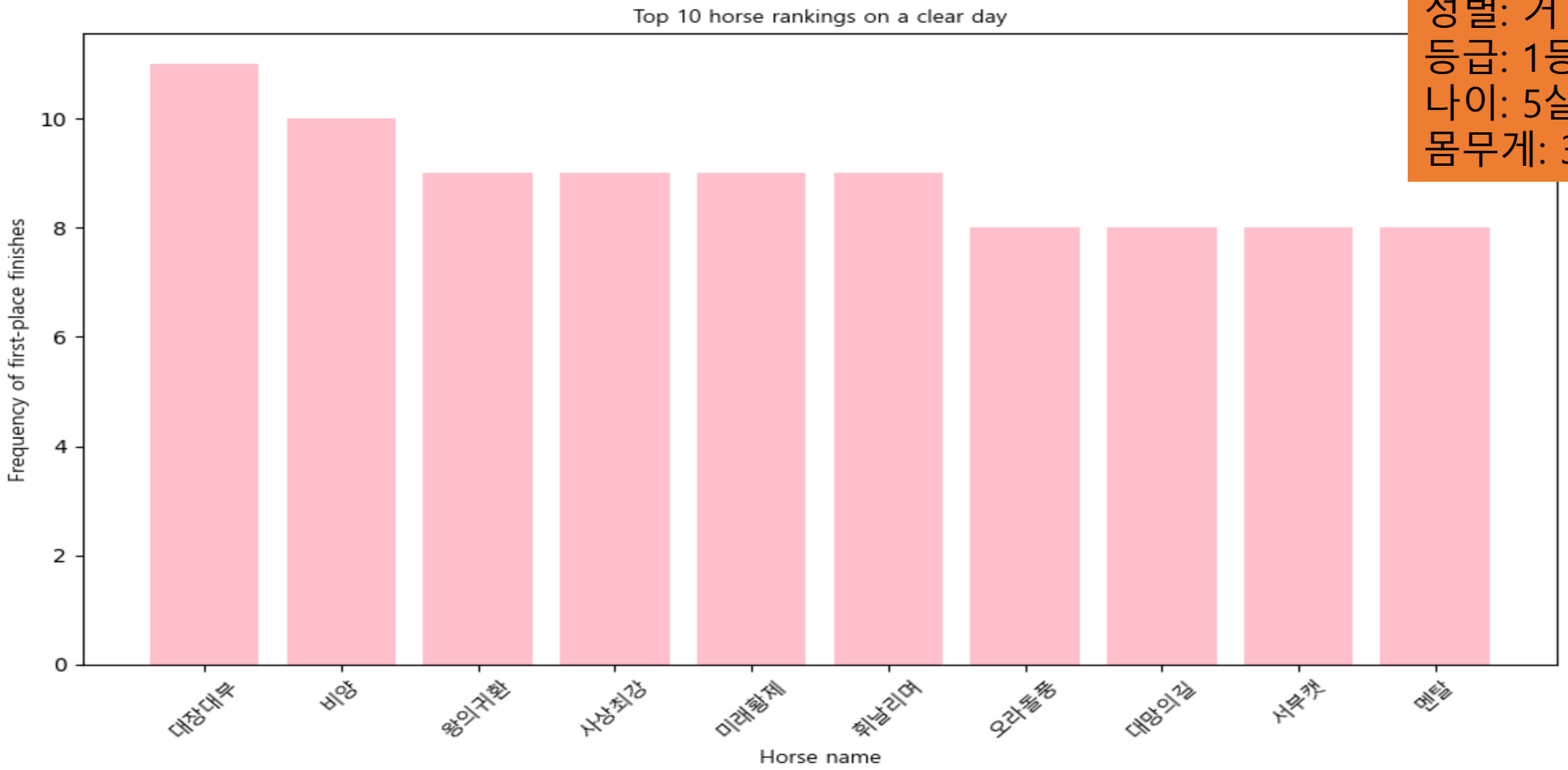
■ 거리 별 1등을 가장 많이 한 말의 이름 (1200m)



이름: 대장대부
성별: 거
등급: 1등급
나이: 5살
몸무게: 313.5kg

03 프로젝트 수행 절차 및 과정

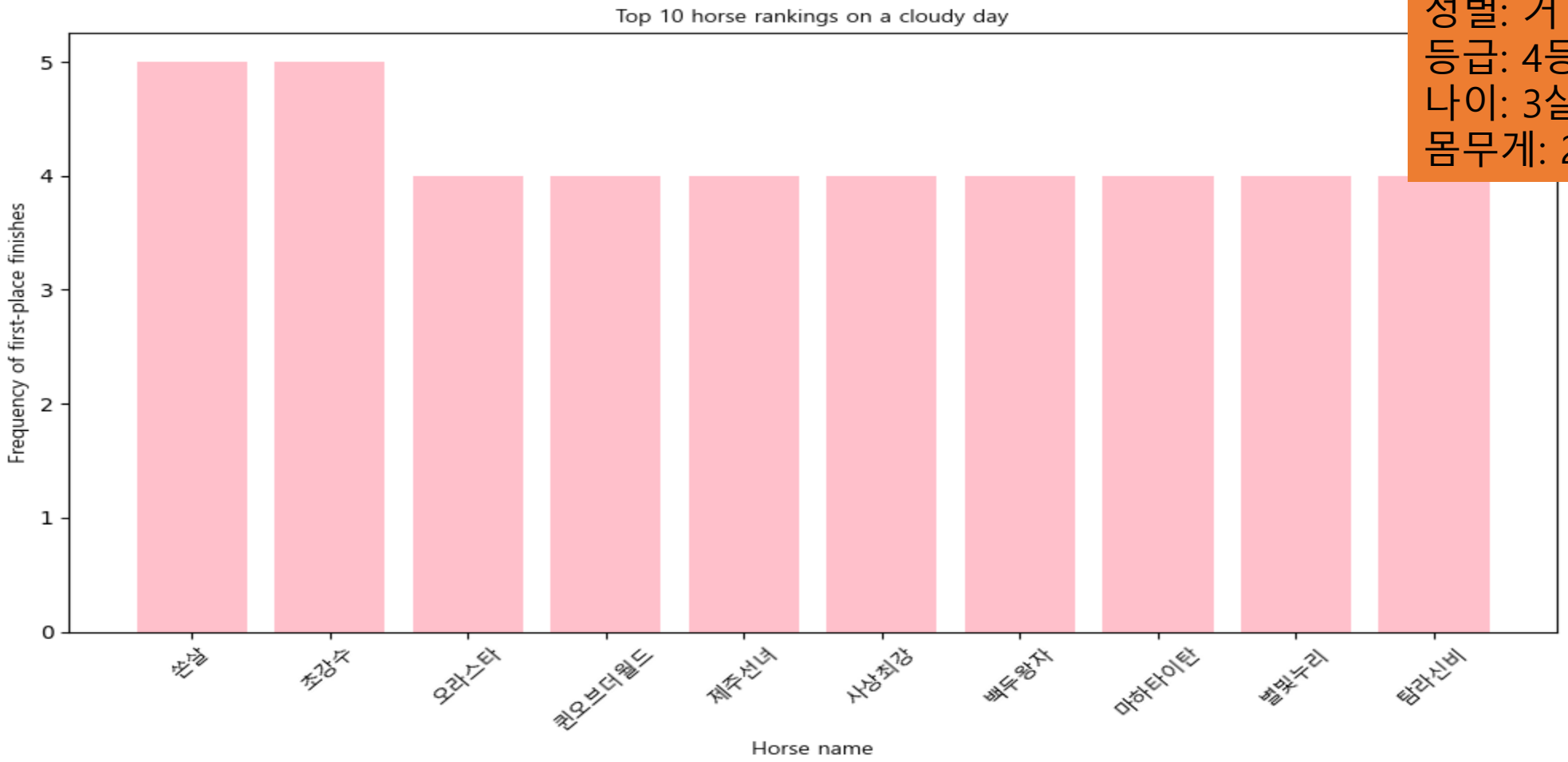
■ 날씨 별 1등을 가장 많이 한 말의 이름 (맑음)



이름: 대장대부
성별: 거
등급: 1등급
나이: 5살
몸무게: 313.5kg

03 프로젝트 수행 절차 및 과정

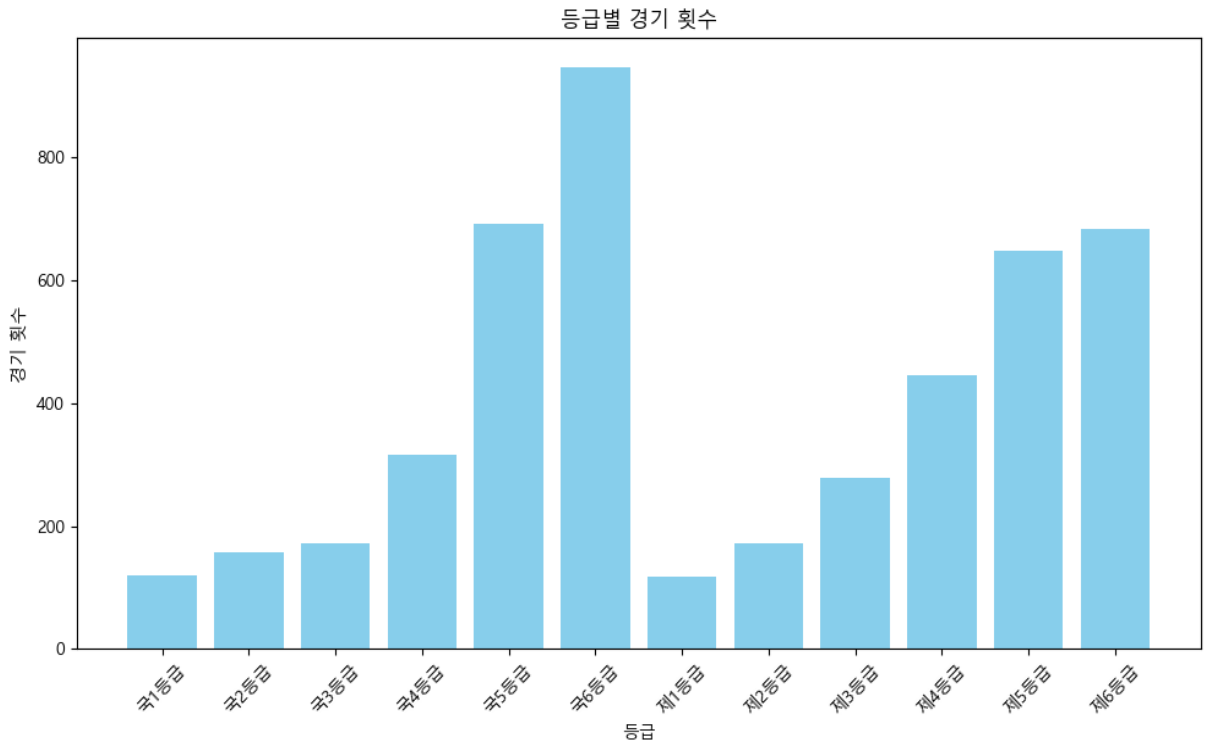
■ 날씨 별 1등을 가장 많이 한 말의 이름 (흐림)

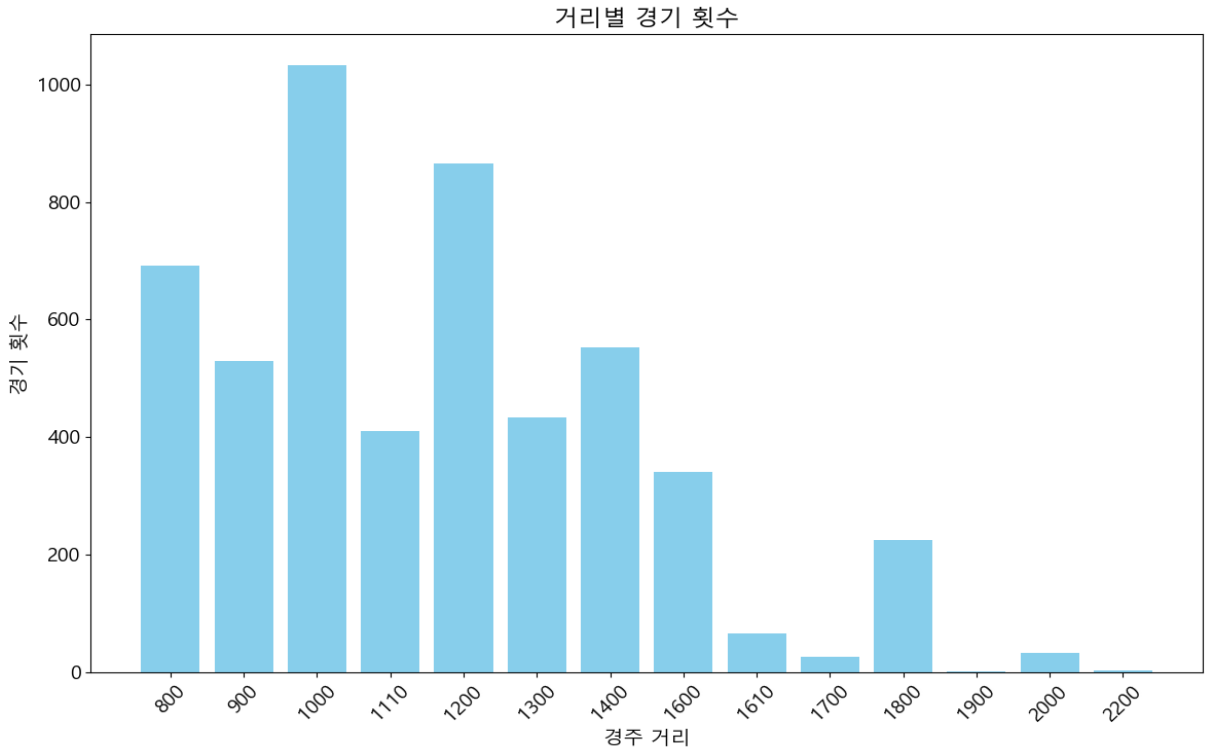


이름: 초강수
성별: 거
등급: 4등급
나이: 3살
몸무게: 284.8kg

03 프로젝트 수행 절차 및 과정

■ 경주 거리 및 경주 등급에 따른 경기 횟수

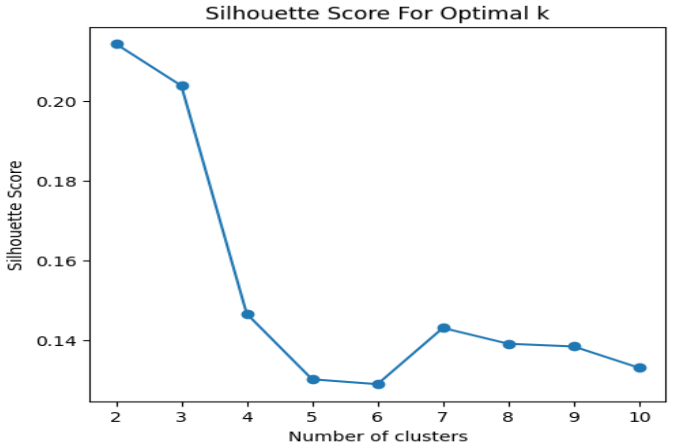


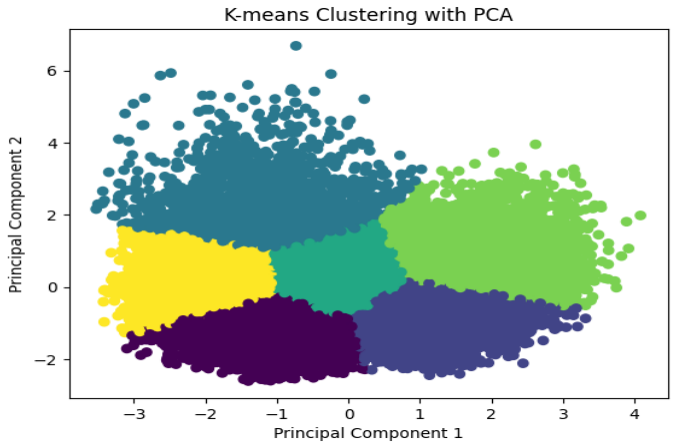


- 경주 거리에 따른 경주 횟수는 1000, 1200, 800 순으로 많았다.
- 경주 등급에 따른 경주 횟수는 등급이 1등급에 가까워 질 수록 적어지는 경향을 보였다.

03 프로젝트 수행 절차 및 과정

■ K-means 및 PCA 분석





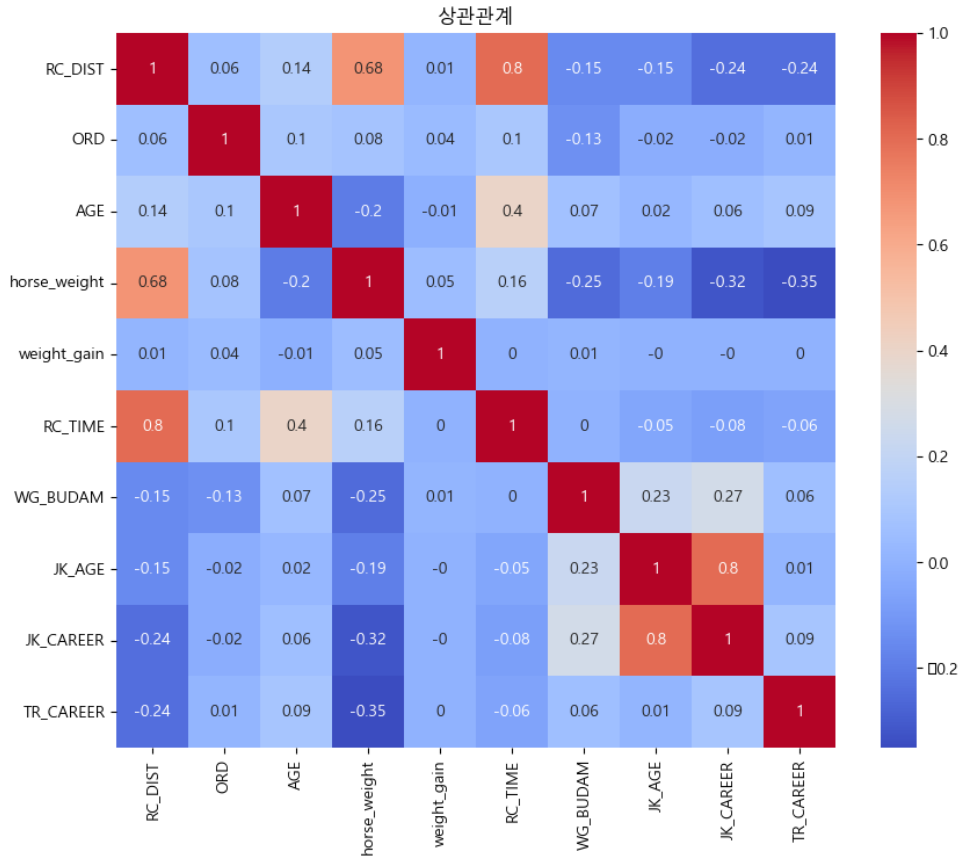

```
Principal Component 1 top contributing features:
AGE: 0.600345354887521
RC_DIST: 0.525259770290738
weight_gain: 0.3752108003365123
WG_BUDAM: 0.3405880049082862
horse_weight: 0.30439042523322474

Principal Component 2 top contributing features:
RANK: 0.7347691869571091
JK_CAREER: 0.5799606989160907
RC_DIST: 0.2858584990503479
WG_BUDAM: 0.18807961636506682
weight_gain: 0.05253893419824085
```

cluster	RC_DIST	AGE	horse_weight	weight_gain	WG_BUDAM	JK_CAREER	TR_CAREER	ORD
0	960.999429	3.704740	287.984580	-0.596421	55.937559	17.631829	11.138587	3.018466
1	1238.687898	3.551752	461.236943	-4.176592	53.487102	9.420382	14.773726	4.118631
2	948.532575	3.804055	281.607368	-0.446038	54.983743	15.097787	25.920757	3.633700
3	1250.671087	3.305003	473.897328	6.062688	54.224734	11.460117	14.742616	2.726743
4	1670.466837	4.493622	481.655612	-0.200510	54.363138	10.283418	14.879082	3.525255
5	1155.399198	7.836675	291.719285	-0.248633	56.358549	15.986876	20.740066	3.967189

04 프로젝트 수행 결과

■ 상관 관계 분석



○ 특별한 상관관계를 발견하지 못함

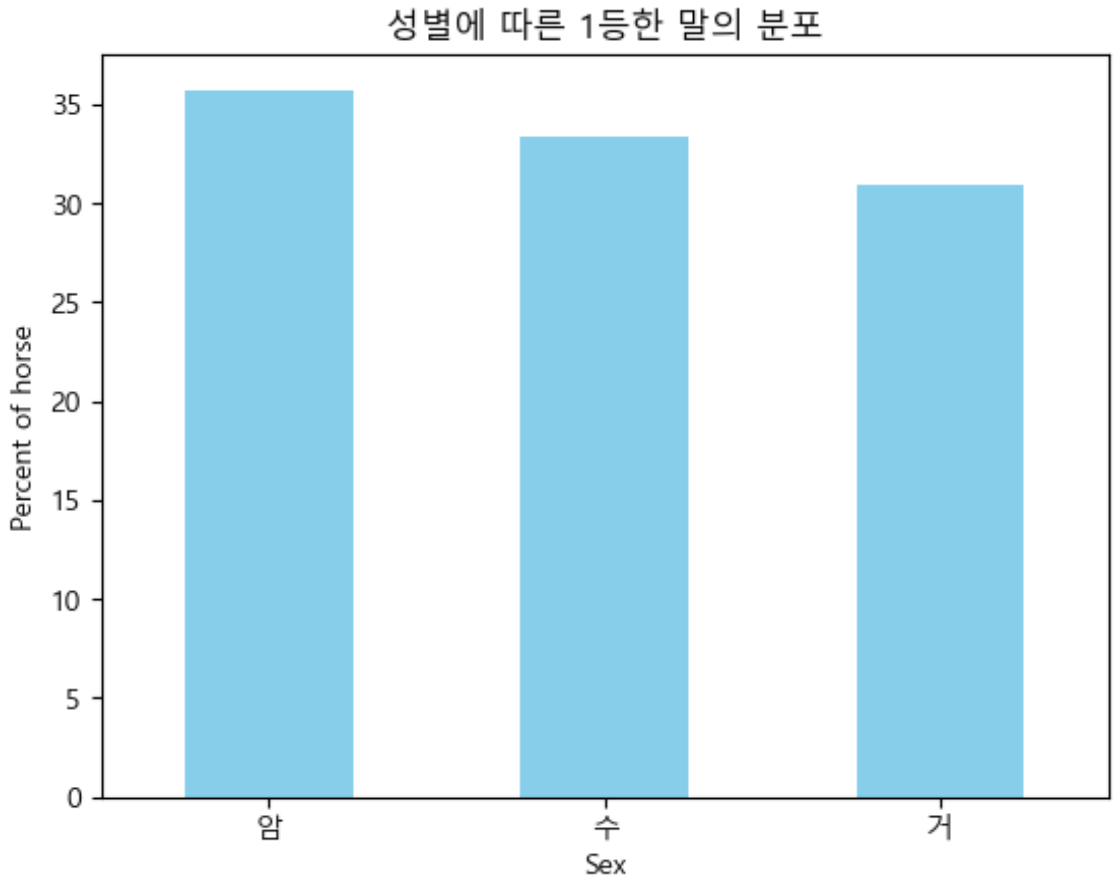
→ 데이터가 정규분포의 모양인 게 원인으로 생각

○ 순위와 그나마 상관관계가 높은 데이터를 선택

- 나이, 기수의 무게, 말의 체중, 성별

03 프로젝트 수행 절차 및 과정

■ 성별에 따른 1등 분포



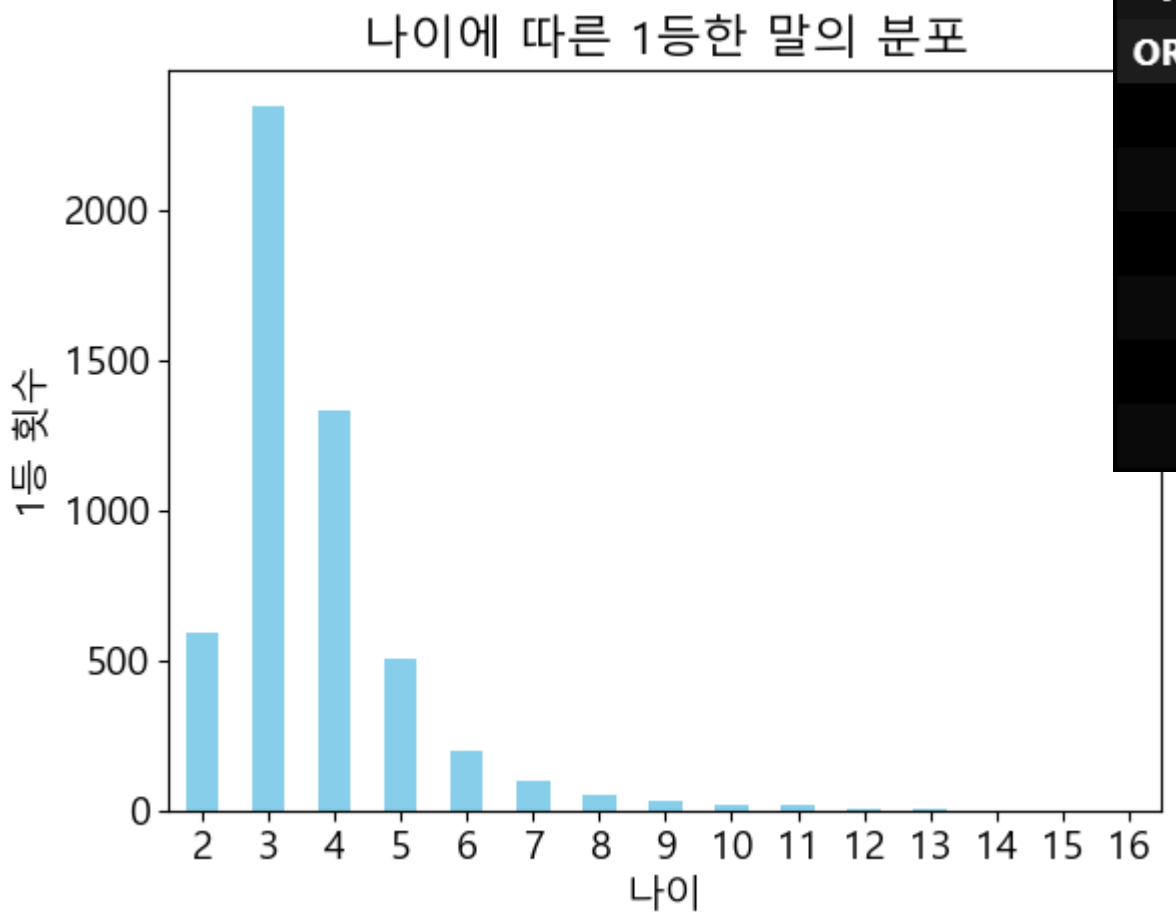
SEX	거	수	암
ORD			
1	1611	1739	1865
2	1679	1670	1869
3	1750	1454	2003
4	1777	1444	1989
5	1770	1450	1978
6	1810	1359	2045

p-value: 1.9498677734854782e-46

- 성별에 따른 승률은 암컷이 높게 나타났지만 카이제곱검정 결과 수컷의 승률이 높게 나타났음.

04 프로젝트 수행 결과

■ 나이에 따른 1등한 횟수

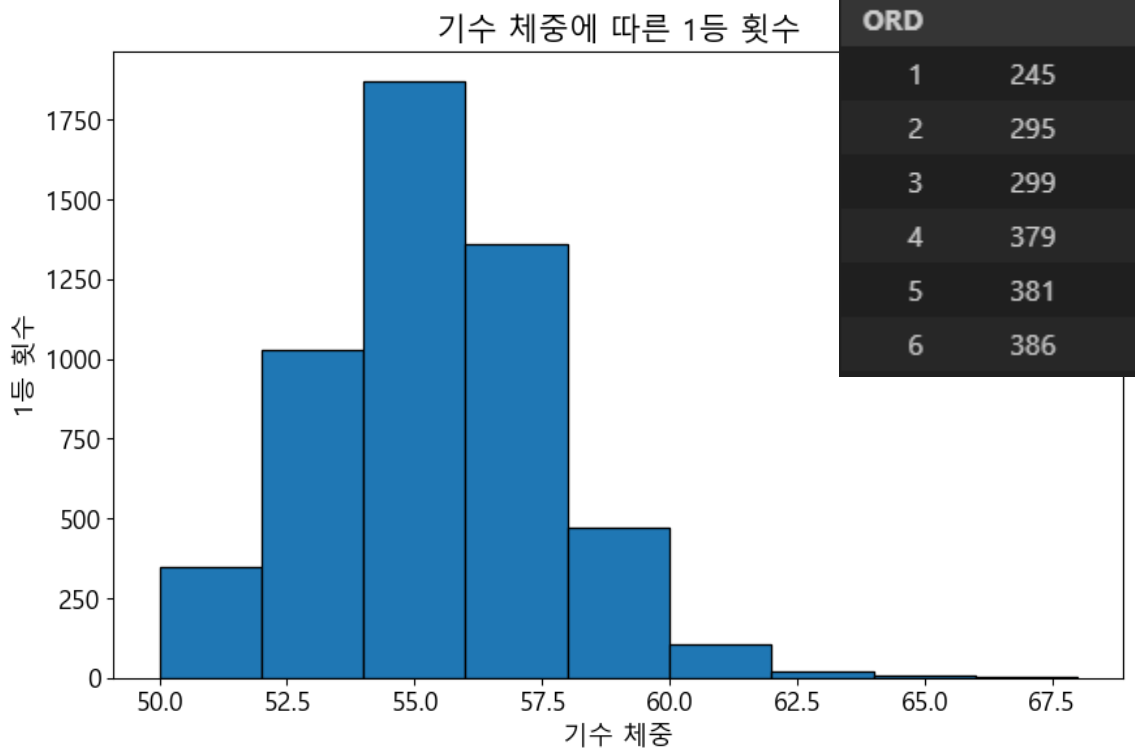


AC / ORD	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 6]	(6, 7]	(7, 8]	(8, 9]	(9, 10]
1	591	2349	1336	505	199	96	53	33	19
2	435	2040	1429	662	289	159	78	50	27
3	372	1852	1472	790	325	170	98	47	31
4	351	1746	1446	834	388	198	119	44	37
5	354	1672	1457	837	385	215	121	67	45
6	312	1595	1467	857	476	234	137	60	34

p-value: 8.190511555092293e-225

04 프로젝트 수행 결과

■ 기수 체중에 따른 1등 횟수



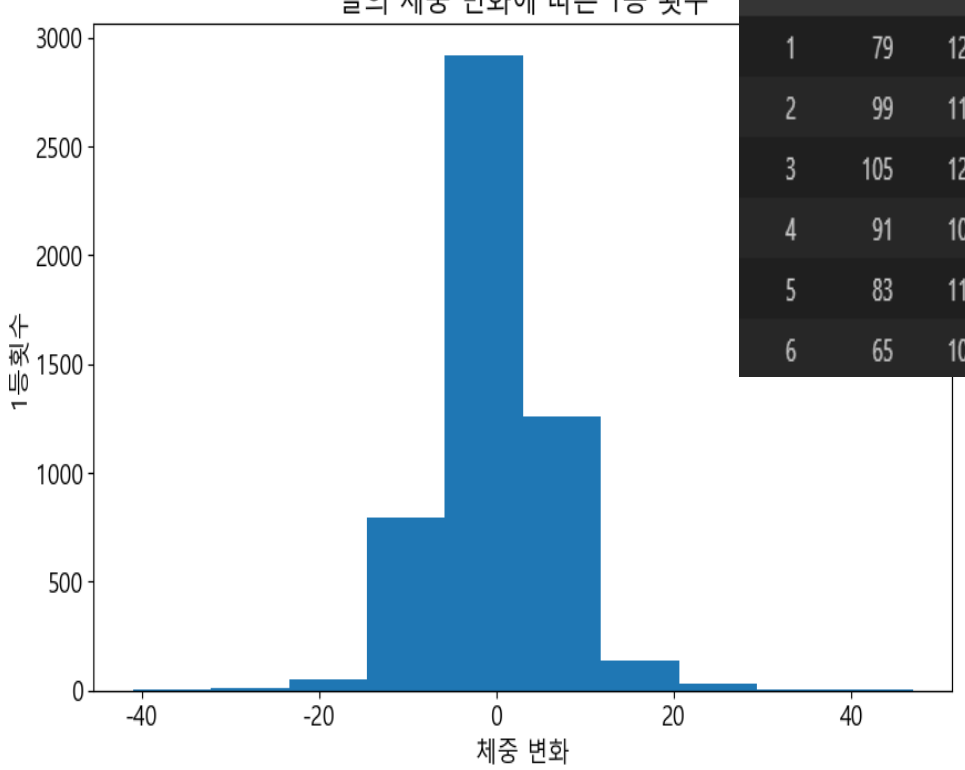
WGB	(50, 51]	(51, 52]	(52, 53]	(53, 54]	(54, 55]	(55, 56]	(56, 57]	(57, 58]	(58, 59]	(59, 60]
ORD										
1	245	407	512	946	876	914	586	383	148	127
2	295	445	546	894	899	861	589	375	126	115
3	299	498	543	931	877	788	571	373	138	112
4	379	517	542	901	839	810	572	338	129	97
5	381	567	531	971	858	785	487	321	114	114
6	386	593	557	981	860	726	515	309	107	94

p-value: 7.405102284692306e-149

- 기수 체중에 따른 1등 횟수는 55kg가 많았지만 카이제곱 검정 결과 1등 확률은 55.7-57.6kg의 기수에서 높게 나타났다.

04 프로젝트 수행 결과

■ 체중 변화에 따른 1등 횟수

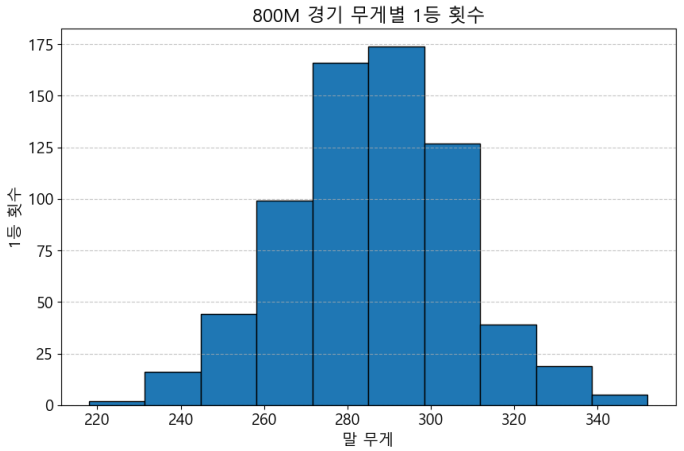


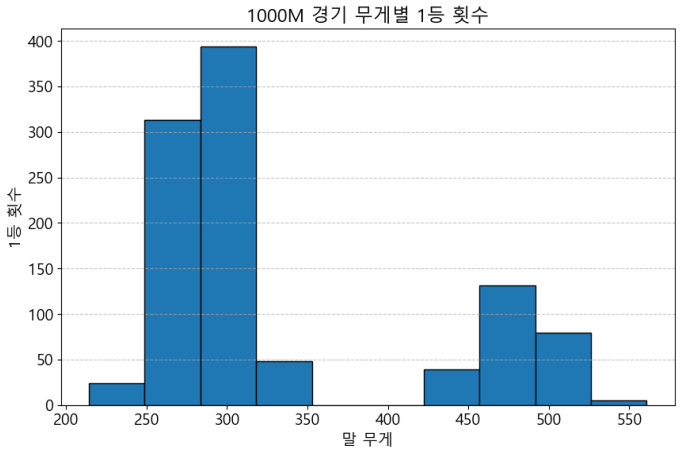
WGC	(-10, -9]	(-9, -8]	(-8, -7]	(-7, -6]	(-6, -5]	(-5, -4]	(-4, -3]	(-3, -2]	(-2, -1]	(-1, 0]	(0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 6]	(6, 7]	(7, 8]	(8, 9]
ORD																			
1	79	125	146	210	300	331	334	397	393	383	380	402	274	244	165	155	102	113	89
2	99	116	140	184	226	318	344	421	433	439	392	346	275	266	198	160	116	99	76
3	105	129	138	199	223	288	322	454	447	429	377	373	309	251	203	166	110	101	82
4	91	104	147	171	261	295	339	415	387	424	409	395	296	272	193	180	121	95	61
5	83	112	132	202	222	288	344	402	407	422	408	388	298	225	199	164	140	114	72
6	65	107	154	177	254	301	360	420	411	394	404	400	284	239	201	168	132	92	84

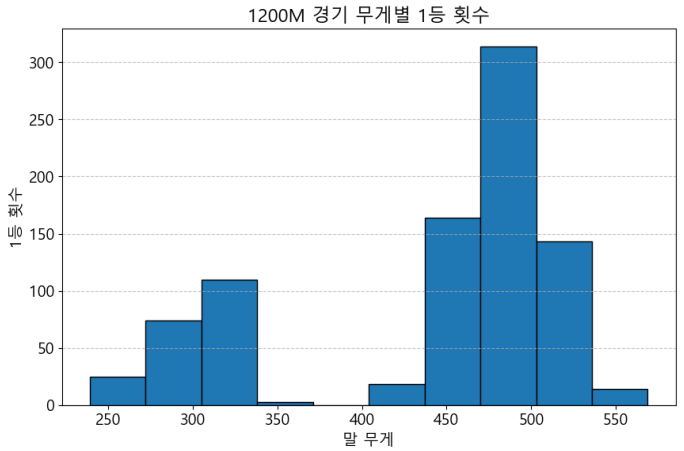
p-value: 1.302588825692841e-05

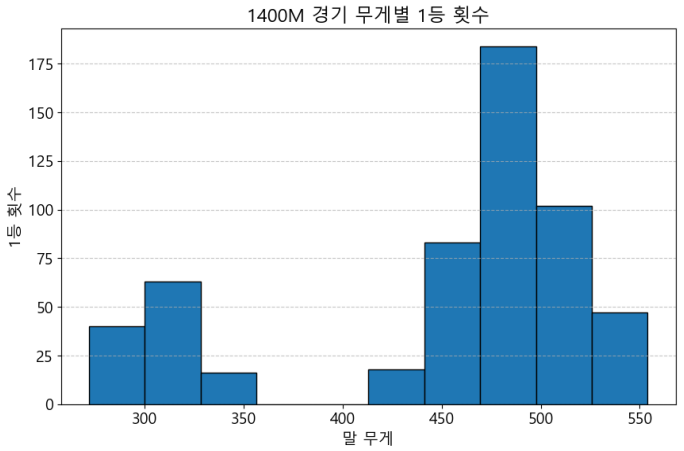
03 프로젝트 수행 절차 및 과정

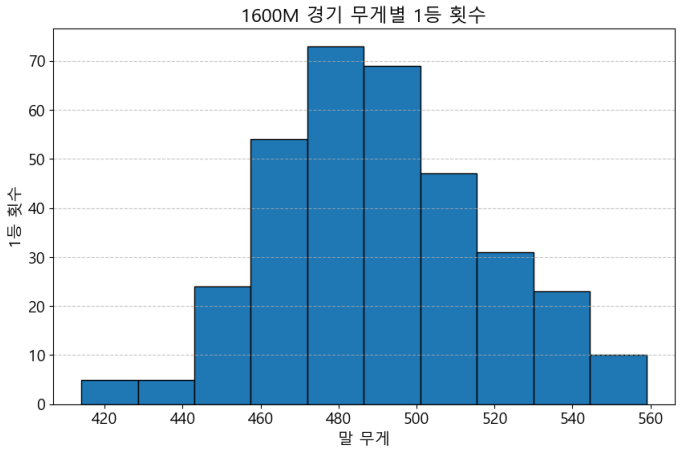
■ 거리 및 체중에 따른 1등의 분포

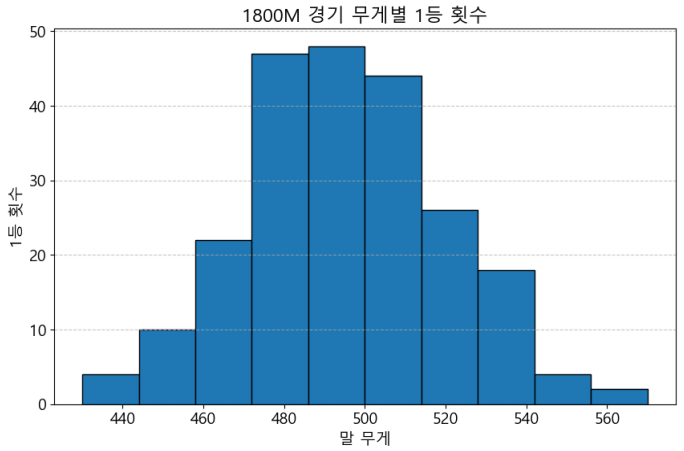






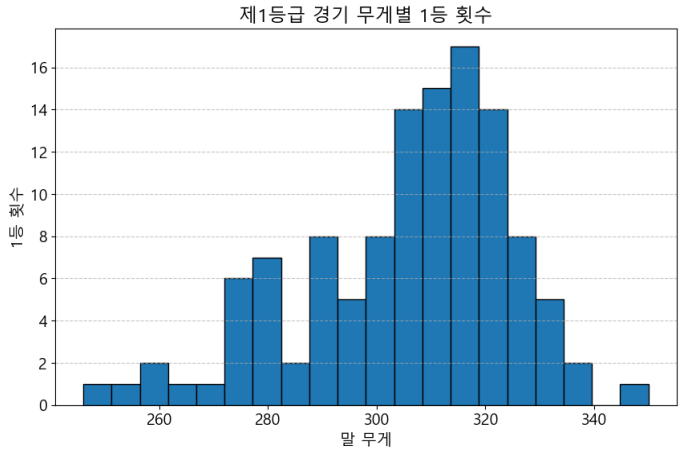


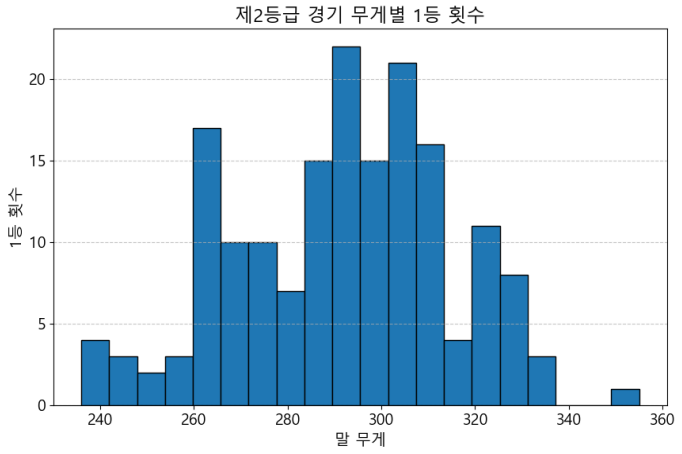


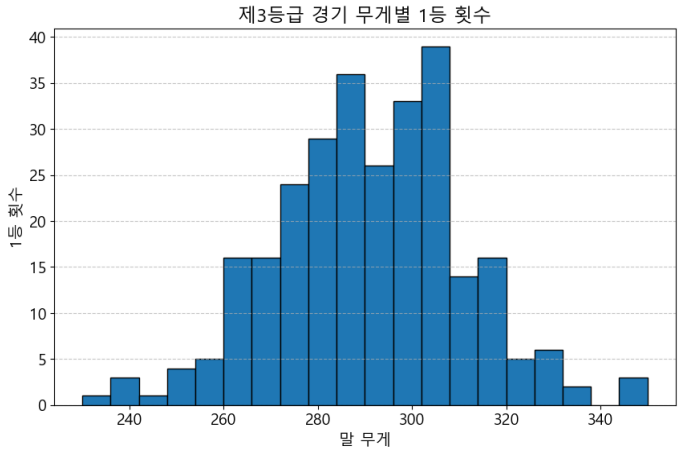


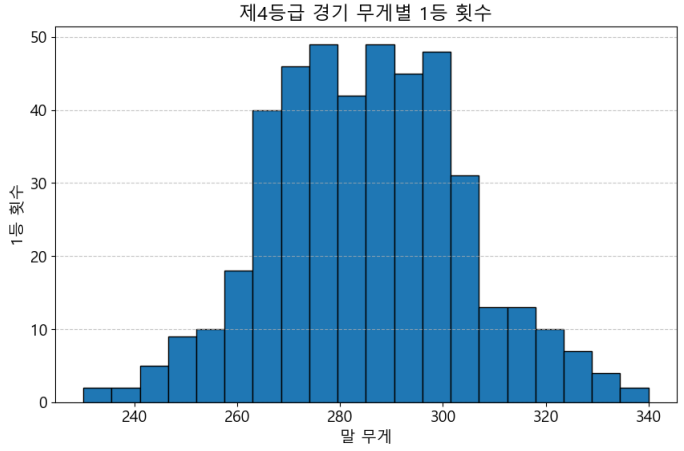
03 프로젝트 수행 절차 및 과정

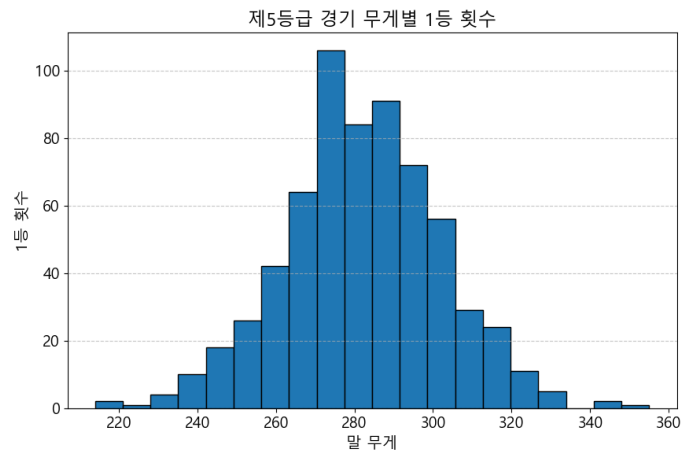
■ 등급 및 체중에 따른 1등의 분포

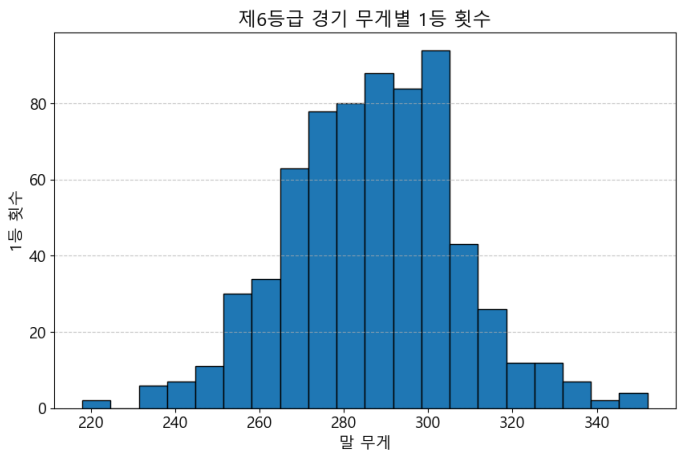












03 프로젝트 수행 절차 및 과정

■ 거리 및 체중에 따른 1등의 분포

horse_weight_bins	[250, 260)	[260, 270)	[270, 280)	[280, 290)	[290, 300)	[300, 310)	[310, 320)	[320, 330)	[330, 340)	[340, 350)	[350, 360)
ORD											
1	123	263	397	478	467	432	290	193	72	25	4
2	165	296	445	501	478	341	234	130	62	27	1
3	225	300	458	497	464	301	221	111	61	16	2
4	193	305	493	489	458	327	200	107	38	5	1
5	236	333	479	436	405	326	214	97	68	25	1
6	203	345	489	499	441	289	201	99	40	22	1

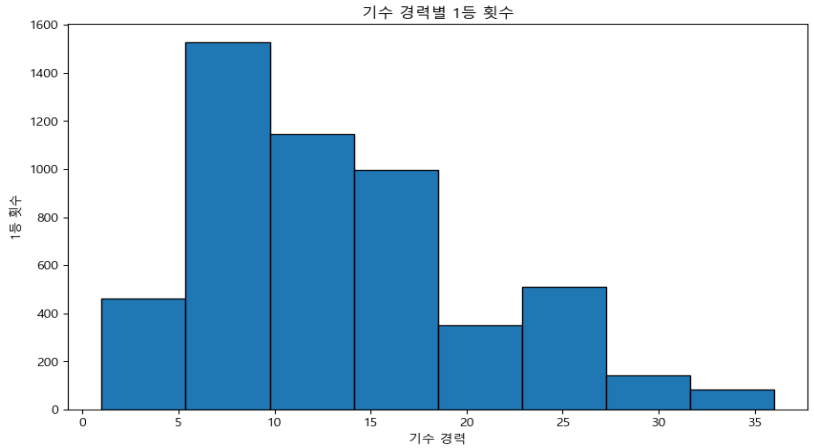
p-value: 1.380872257385084e-33

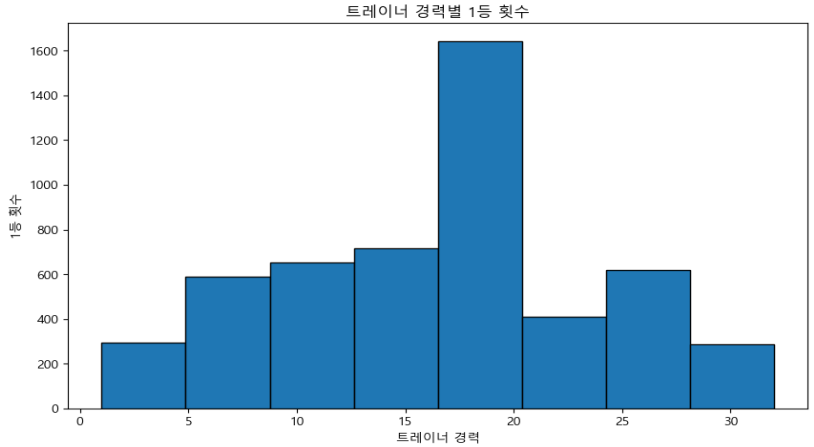
horse_weight_bins	[400, 410)	[410, 420)	[420, 430)	[430, 440)	[440, 450)	[450, 460)	[460, 470)	[470, 480)	[480, 490)	[490, 500)	[500, 510)	[510, 520)	[520, 530)	[530, 540)	[540, 550)
ORD															
1	1	12	23	51	77	177	274	360	372	292	283	187	115	98	34
2	0	7	28	67	106	205	308	333	369	315	250	153	128	64	30
3	4	17	42	64	150	217	281	324	369	295	245	166	104	54	29
4	8	19	46	89	139	236	277	343	351	275	222	167	123	57	19
5	7	22	54	109	130	236	307	324	353	286	216	146	106	54	14
6	9	20	59	96	137	233	303	351	351	276	223	141	100	47	21

p-value: 2.356848304914838e-22

04 프로젝트 수행 결과

■ 기수 및 트레이너의 경력에 따른 1등 횟수





JCC ORD	(-0.001, 5.0]	(5.0, 10.0]	(10.0, 15.0]	(15.0, 20.0]	(20.0, 25.0]	(25.0, 30.0]
1	461	1855	1056	1091	380	147
2	540	1774	1092	1015	359	145
3	571	1760	1043	984	354	149
4	611	1698	1048	1005	326	166
5	627	1700	1026	1012	292	175
6	699	1623	1001	972	323	159

p-value: 4.24471588758243e-105

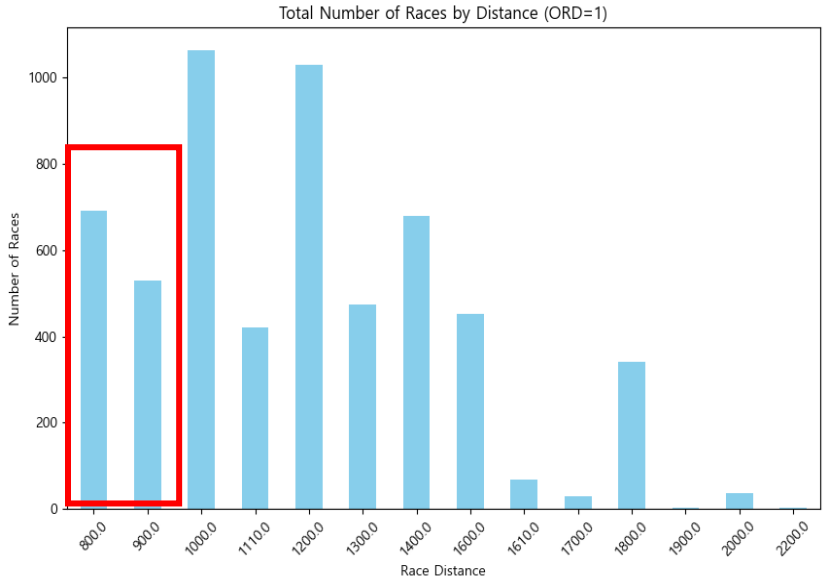
TRC ORD	(-0.001, 5.0]	(5.0, 10.0]	(10.0, 15.0]	(15.0, 20.0]	(20.0, 25.0]	(25.0, 30.0]
1	540	586	1103	1669	410	709
2	519	535	1028	1757	382	772
3	488	523	926	1756	413	818
4	459	497	966	1800	384	795
5	459	459	988	1738	442	791
6	440	481	879	1845	417	816

p-value: 3.930171145438816e-160

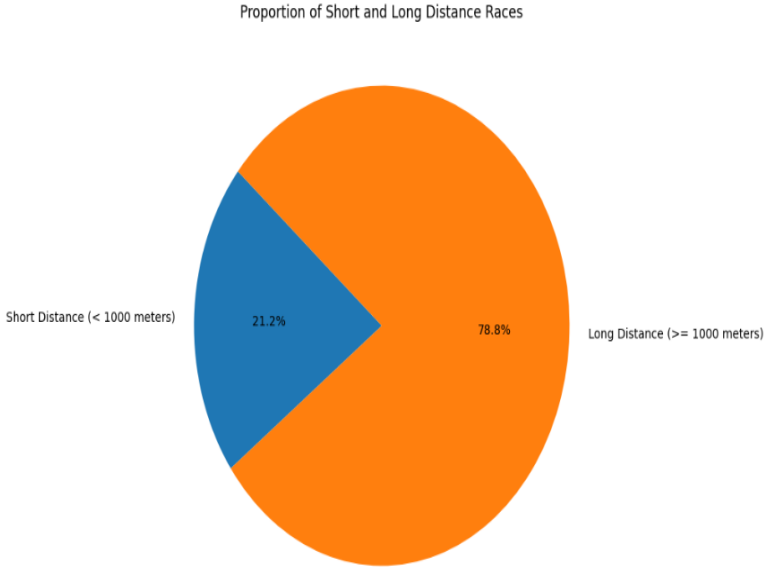
03 프로젝트 수행 절차 및 과정

단거리 + 흐린날씨에 성적이 좋은 말들의 특성 조사

전체 경기의 거리를 구하기



20%정도로 데이터 축약



줄인 이유: 유효한 특성을 찾아보기 위해 최소한의 데이터를 이용하고자 함.

03 프로젝트 수행 절차 및 과정

	WEATHER	NAME	First_Place_Count
0	눈	한국	21
1	맑음	한국	856
2	비	한국	70
3	안개	한국	29
4	흐림	한국	246

흐림날씨를 고른 이유:
데이터를 줄여서 유의미한 특성을 쉽게 찾기 위해서

흐림날씨 상위3 마리의 날씨별 성적

	Weather	Win_Count
0	흐림	10
1	맑음	6
2	비	1
3	안개	1

맑은 날씨의 상위 top 10

	HR_NAME	First_Place_Count
39	광아의명성	4
44	광해왕자	4
552	푸른장군	4
1	가왕신화	3
10	강한보스	3
15	거서간	3
20	경성의별	3
31	광명시대	3
40	광아의여걸	3
52	금누리	3

흐림 날씨의 상위 top 10

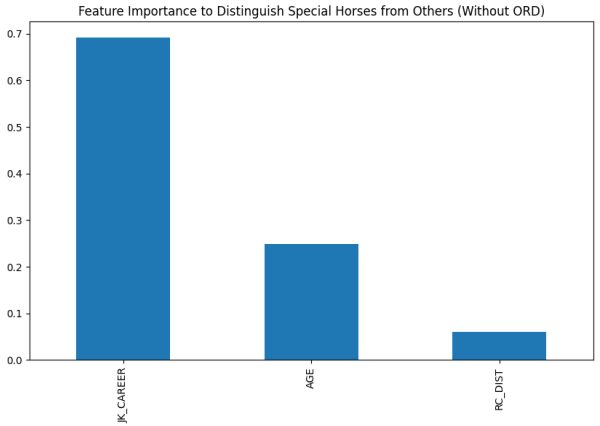
	HR_NAME	First_Place_Count
171	초강수	4
76	별빛누리	3
121	오라여걸	3
27	녹색공원	2
28	녹색비상	2
36	달의강자	2
51	명성스타	2
68	백두왕자	2
69	백두정복	2
86	사상최강	2

경기수의 차이를 보면 흐림날씨의 상위 3마리는 유의미한 승률을 보이는 것으로 판단됨.

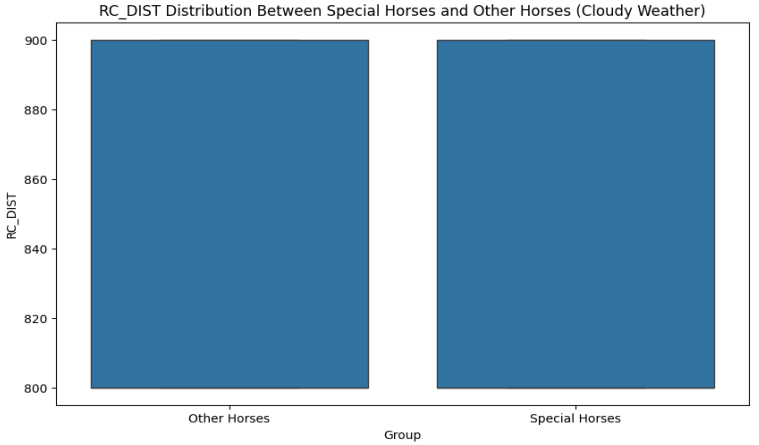
03 프로젝트 수행 절차 및 과정

랜덤 포레스트 모델을 사용해서 특성 중요도를 분석

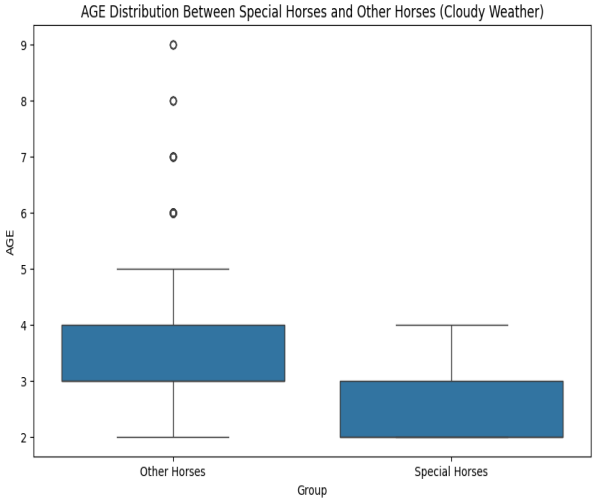
전체결과



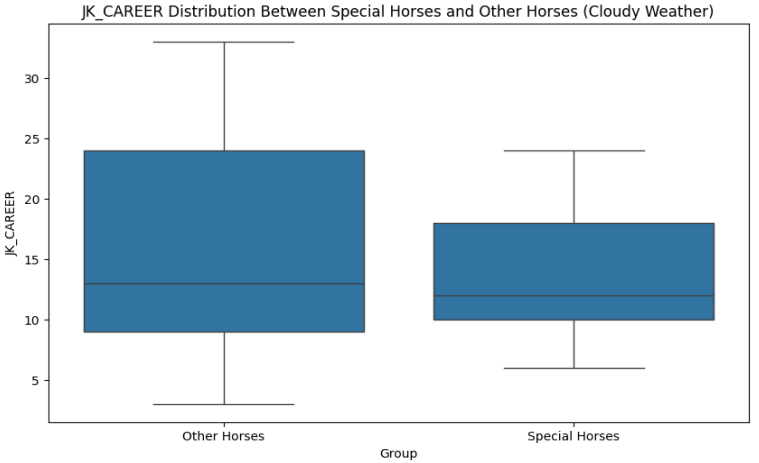
경기장 거리



말의 나이



기수의 경력



최종 결론:

- **기수의 경력**과 **말의 나이**가 흐린 날씨에서 성과에 큰 영향을 미칠 수 있다.

- **경력이 많은 기수**와 **나이가 어린 말**들이 흐린 날씨에서 승리할 가능성이 높을 수 있다.

- **나이가 어리고 숫컷인 말이 높은 승률을 기록하였다.**

05 자체 평가 의견

오용석

- 데이터 분석 과정에 들어와서 처음 해보는 프로젝트로 1등을 맞추는 데 너무 집중한 나머지 다른 등수 또한 같은 결과가 나올 수 있다는 점을 모르고 있었다. 하지만 이번 경험을 통해서 데이터 분석이란 결국 내가 분석할 데이터의 모든 내용을 머릿속에 그리고 있어야 하고 다양한 방식으로 접근해야 한다는 것을 알게 되었다.

이정인

- 데이터 분석을 배우면서 생각보다 어렵지 않다고 생각해서 편한 마음으로 임했는데, 시간이 갈수록 자꾸 오류가 생기기도 하고 원하는 방향이 있는데 어떤 코드를 써야 하고 어떤 위치에 써야 할지 몰라서 생각보다 연습과 전체적으로 크게 보는 연습을 해야 할 거 같다.

원석재

- 방대한 데이터를 분석하는 기초적인 방법을 이해할 수 있었고, 코드 분석 및 활용 능력을 더욱 향상 시킬 필요성을 느꼈다. ChatGPT를 효과적으로 활용할 수 있을 정도로 분석 능력을 키우고자 한다.