Hong Kong Horse Racing Acquisition and Processing

Jinru KUANG, Zhengyi LIU,

Xiaqing WU, Chenyan LYU

Department of Media and Communication

COM5507 Social Media Data Acquisition and Processing

Xiaofan LIU

December 13, 2024

Author List and Contributions

Members	Contributions
Jinru KUANG	25%
Zhengyi LIU	25%
Xiaqing, WU	25%
Chenyan LYU	25%
Total	100%

Abstract

This report presents an initial exploration into the acquisition, processing, and preliminary analysis of Hong Kong horse racing data, setting the stage for more advanced predictive modeling in the future. Two primary data sources were selected to ensure both reliability and richness: the Hong Kong Jockey Club's official website, which offers authoritative race records, historical datasets, and detailed horse and jockey profiles, and MaJing, a local platform that provides supplementary insights and reference materials. Data extraction and preprocessing were carried out using Python, encompassing steps such as data cleaning, normalization, and restructuring, to ensure consistency and usability. Subsequently, fundamental visualization techniques—including histograms, scatter plots—were employed to uncover key patterns within the data. While the processed datasets and initial visualizations offer a useful starting point for newcomers to Hong Kong horse racing, it is important to note that this report represents only the initial stage of a more comprehensive research effort. Future work will focus on constructing and refining predictive models to facilitate more accurate forecasts and data-driven decision-making in the domain of Hong Kong horse racing.

Introduction

Background and Motivation

Horse racing, as a unique social event, embodies a vibrant aspect of Hong Kong's rich cultural tapestry. This exhilarating spectacle is not just about the speed of the horses or the strategy of betting, it's an occasion that brings together people from diverse backgrounds, creating a shared experience filled with excitement, anticipation, and community spirit. Both tourists and local residents flock to the races, drawn by the electrifying atmosphere that permeates the racecourse. The sights and sounds—the thundering hooves, the cheering crowds, and the elegant fashion—combine to create a memorable experience that captures the essence of Hong Kong's dynamic culture.

The motivation behind this project stems from the desire to help horse racing enthusiasts make more informed judgments regarding their betting strategies. By undertaking a comprehensive process that includes data scraping, selection, cleaning, analysis, and visualization, we aim to provide valuable insights into horse racing in Hong Kong. Our goal is to present clear and relevant information that aids bettors and fans in navigating the complexities of the sport. Through this meticulous approach, we aspire to create a resource that

empowers participants in this exhilarating arena, ultimately enhancing their overall experience and confidence in their betting choices.

Others Work and Our Angle

Before delving deeper into our research, it is worth mentioning that there have been some excellent studies conducted in this field. Although most of the studies utilize machine learning algorithms to predict horse racing outcomes, we can still draw inspiration from the data types they employ to enhance the efficiency and accuracy of our data collection process. Gupta and Singh (2023) explored the impact of multiple factors on horse racing outcomes in their study, including information on horse characteristics and historical performance, to identify key variables that can improve prediction accuracy. Chavan et al. (2023) analyzed horse race performance and other related information, focusing on identifying the main factors that affect performance and providing relevant data analysis results. Both studies utilized horse racing data, with a focus on analyzing the impact of historical race records on predicting horse racing outcomes.

In contrast to these two articles from India, our angle is to leverage the insights from various studies to conduct research specific to horse racing in Hong Kong, where existing literature is

relatively scarce. Therefore, we are drawing on the data types identified in previous research and concentrating on comprehensive scraping of detailed information, historical race records, and relevant statistical data concerning current horses in Hong Kong.

Methods

Data Source

Regarding data source selection, we chose the Hong Kong Jockey Club and Majing website as the authoritative data source for our research. The Hong Kong Jockey Club provides up-to-date, official information, including basic information about the horses, race records, and relevant background information. At the same time, the rich historical data on *Majing* also provides us with a deep perspective, enabling us to analyze recent and long-term competition trends.

The URLs we scraped are as follows:

https://racing.hkjc.com/racing/information/chinese/Horse/SelectHorse.aspx

https://racing.on.cc/cgi-bin/srh/search/search.cgi

W1-Hong Kong Jockey Club

On the Hong Kong Jockey Club website, we scraped the basic information and recent 3 season records of all active horses in Hong Kong based on the word count of their names.

Targeted web pages are as follows:

賽事資料(本地) - 馬匹資料 - 馬匹資料

返回馬匹搜尋頁

馬名字數			
• 未命名馬匹	• 二字馬	• 三字馬	• 四字馬

二字馬 (依評分由大至小橫向排列)									
• 安遇	(104)	• 安騁	(104)	• 氣勢	(98)	• 增有	(89)	• 球星	(83)
• 瑪瑙	(82)	• 耀寶	(78)	• 晴王	(77)	• 玩笑	(76)	• 米奇	(76)
• 優才	(73)	• 十力	(72)	• 伊臣	(71)	• 馬力	(71)	• 祝願	(71)
• 豐盛	(71)	• 快路	(70)	• 將義	(70)	• 飲杯	(69)	• 增勁	(69)
• 俊才	(68)	• 銀進	(67)	• 奔放	(67)	• 堅闖	(66)	• 增強	(66)
• 佳登	(66)	• 福星	(66)	• 赤壁	(66)	• 翠紅	(65)	• 星價	(65)
• 三強	(65)	• 快搏	(65)	• 安都	(64)	• 凡喜	(64)	• 旌鷹	(63)
• 論文	(62)	• 初戀	(61)	• 爵登	(61)	• 大才	(61)	• 攻頂	(60)
• 翹峯	(60)	• 勝意	(58)	• 爆笑	(58)	• 銀騰	(57)	• 安泰	(56)
• 帥炸	(56)	• 同喜	(56)	• 爆熱	(56)	• 定數	(55)	• 真感	(54)
• 神馳	(53)	• 上校	(53)	• 摩界	(53)	• 飛雲	(53)	• 黑白	(53)

Selecting horses

安遇 (G236)

- 往績紀錄
- 馬匹評分/體重/名次
- 所跑途程賽績紀錄
- 晨操紀錄
- 傷患紀錄
- 搬遷紀錄
- 海外賽績紀錄
- 血統簡評
- 其他馬匹

*包括本地及海外賽績及獎金

出生地 / 馬齡	: 愛爾蘭 / 5	練馬師	: 賀賢
毛色 / 性別	: 棕 / 閹	馬主	: 王賢訊
進口類別	: 自購馬	現時評分	: 104
今季獎金*	: $187,250	季初評分	: 107
總獎金*	: $14,181,025	父系	: Churchill
冠-亞-季-總出賽次數*	: 7-2-2-28	母系	: Enrol
最近十個賽馬日出賽場數	: 2	外祖父	: Pivotal
現在位置 (到達日期)	: 香港 (24/09/2024)	同父系馬	: 美麗之光 瀏覽
進口日期	: 16/10/2021		

Horse detailed information

馬匹近三季往績紀錄 - 安遇

近三季往績 | 三季前往績 | 所有往績

場次	名次	日期	馬場/跑道/賽道	途程	場地狀況	賽事班次	檔位	評分	練馬師	騎師	頭馬距離	獨贏賠率	實際負磅	沿途走位	完成時間	排位體重	配備	賽事重播
24/25 馬季																		
190	05	17/11/24	沙田草地"B+2"	2000	好	G2	6	104	賀賢	布文	5	34	123	10 10 10 9 5	2.00.51	1174	B	
153	08	03/11/24	沙田草地"C+3"	1800	好/快	G3	8	105	賀賢	霍宏聲	3-1/2	42	121	12 12 12 12 8	1.46.97	1161	B	
117	12	20/10/24	沙田草地"B+2"	1200	好/快	G2	6	107	賀賢	班德禮	7-1/2	183	116	12 13 12	1.08.79	1168	B2/TT-	
23/24 馬季																		
719	07	02/06/24	沙田草地"B"	1600	好/快	G3	4	111	文家良	田泰安	13-3/4	26	127	4 3 6 7	1.36.15	1115	TT	
566	13	07/04/24	沙田草地"B+2"	1600	好	G2	4	113	文家良	田泰安	17-1/4	10	123	7 6 10 13	1.36.79	1125	TT	
452	07	25/02/24	沙田草地"A+3"	2000	好	G1	2	114	文家良	田泰安	5-1/2	16	126	5 6 7 7 7	2.01.20	1146	TT	
395	08	04/02/24	沙田草地"B+2"	1800	好	G3	12	114	文家良	梁家俊	3	15	132	2 3 5 5 8	1.47.89	1155	TT	
330	08	10/01/24	跑馬地草地"B"	1800	好	G3	12	114	文家良	梁家俊	4-1/4	11	135	7 8 8 7 8	1.48.99	1164	TT	
238	08	10/12/23	沙田草地"A"	1600	好	G1	1	114	文家良	梁家俊	4-3/4	24	126	6 6 7 8	1.34.88	1164	TT	
144	01	05/11/23	沙田草地"C+3"	1800	好	G3	6	110	文家良	梁家俊	1-1/4	12	128	3 4 4 4 1	1.47.40	1164	TT	
088	02	15/10/23	沙田草地"A+3"	1600	好	G2	4	105	文家良	梁家俊	1/2	14	117	3 3 4 2	1.34.28	1156	TT	
039	06	24/09/23	沙田草地"C"	1400	好	G3	6	105	文家良	梁家俊	10-1/2	4.4	123	5 6 4 6	1.23.59	1153	TT	
22/23 馬季																		
834	01	16/07/23	沙田草地"A"	1600	好	1	7	100	文家良	梁家俊	頸位	4.9	128	9 9 5 1	1.34.93	1165	TT	
749	01	14/06/23	跑馬地草地"B"	1650	好	2	10	89	文家良	梁家俊	4-3/4	3.2	124	5 4 1 1	1.38.87	1153	TT	
576	01	12/04/23	跑馬地草地"B"	1650	好	2	6	83	文家良	梁家俊	1	5.9	120	4 4 3 1	1.39.08	1138	TT	
509	09	19/03/23	沙田草地"A"	2000	好	4YO	8	83	文家良	蘇兆輝	3-3/4	124	126	2 2 2 2 9	2.03.37	1152	B-/H-/TT	
453	12	26/02/23	沙田草地"A+3"	1800	好/快	4YO	9	83	文家良	蘇兆輝	8-1/2	28	126	7 6 6 11 12	1.47.62	1152	B/H1/TT	
406	02	08/02/23	跑馬地草地"B"	1650	好	3	5	81	文家良	蘇兆輝	頸位	2.2	132	2 2 2 2	1.39.62	1153	B/TT	
328	01	11/01/23	跑馬地草地"B"	1650	好	3	12	74	文家良	蘇兆輝	1	4.2	131	2 2 1 1	1.39.53	1137	B/TT	
250	01	14/12/22	跑馬地草地"B"	1650	好	3	3	67	文家良	蘇兆輝	2-1/4	3.5	126	2 3 4 1	1.39.13	1129	B1/TT	
166	03	12/11/22	沙田全天候	1650	好	3	6	68	文家良	潘明輝	9	11	121	2 2 2 3	1.38.22	1133	TT	
116	05	23/10/22	沙田草地"B+2"	1600	好/快	3	11	69	文家良	潘頓	2	5.8	124	2 2 2 5	1.33.80	1136	TT	
044	06	25/09/22	沙田草地"C"	1600	好/快	3	2	69	文家良	潘明輝	2-3/4	4.2	124	3 5 5 6	1.35.23	1143	TT	
010	03	11/09/22	沙田草地"A"	1400	好	3	12	69	文家良	潘明輝	1-1/4	8.4	122	4 2 2 3	1.22.58	1140	TT	

Recent 3 seasons racing records

The inspecting the element windows are as follows:

```html
<table width="760" border="0" class=bigborder border="0" cellspacing="1" cellpadding=0>
    <tr>
        <td class=subheader>馬名字數</td>
    </tr>
    <tr>
        <td>
            <table width="760" border="0" border=0 cellspacing=1 cellpadding=0>
                <tr>
                    <td bgcolor="FFFFFF" width="190" class=table_text_two>
                        <ul><li class=table_text_two><a href="/racing/information/chinese/Horse/SelectHorsebyChar.aspx?ordertype=0" class=table_text_two>未命名馬匹</a></li></ul>
                    </td>
                    <td bgcolor="FFFFFF" width="190" class=table_text_two>
                        <ul><li class=table_text_two><a href="/racing/information/chinese/Horse/SelectHorsebyChar.aspx?ordertype=2" class=table_text_two>二字馬</a></li></ul>
                    </td>
                    <td bgcolor="FFFFFF" width="190" class=table_text_two>
                        <ul><li class=table_text_two><a href="/racing/information/chinese/Horse/SelectHorsebyChar.aspx?ordertype=3" class=table_text_two>三字馬</a></li></ul>
                    </td>
                    <td bgcolor="FFFFFF" width="190" class=table_text_two>
                        <ul><li class=table_text_two><a href="/racing/information/chinese/Horse/SelectHorsebyChar.aspx?ordertype=4" class=table_text_two>四字馬</a></li></ul>
                    </td>
                </tr>
            </table>
        </td>
    </tr>
</table>
<br />
<table width="760" border="0" class=bigborder border="0" cellspacing="1" cellpadding=0>
```

The character of the horse's name

```html
<table width="760" border="0" class=bigborder border="0" cellspacing="1" cellpadding=0>
    <tr>
        <td class=subheader>二字馬 (依評分由大至小横向排列)</td>
    </tr>
    <tr>
        <td>
            <table width="760" border=0 cellspacing=1 cellpadding=0>
<tr>
                    <td width=20% bgcolor=FFFFFF class=table_eng_text>
                    <table width="100%" cellspacing="0" cellpadding="0" border=0>
                        <tr>
                            <td class=table_text_two bgcolor=#ffffff align=left style="padding:4px 2px; width:100px;">
                                <ul><li class=table_text_two style="padding:4px 2px"><a href="/racing/information/chinese/Horse/Horse.aspx?HorseId=HK_2021_G236" class=ltable_text>安遇</a></li></ul>
                            </td>
                            <td align="right" class=table_eng_text style="width: 52px; padding: 4px 3px; vertical-align: middle;">(104)</td>
                        </tr>
                    </table>
                </td>
                <td width=20% bgcolor=FFFFFF class=table_eng_text>
                    <table width="100%" cellspacing="0" cellpadding="0" border=0>
                        <tr>
                            <td class=table_text_two bgcolor=#ffffff align=left style="padding:4px 2px; width:100px;">
                                <ul><li class=table_text_two style="padding:4px 2px"><a href="/racing/information/chinese/Horse/Horse.aspx?HorseId=HK_2022_H452" class=ltable_text>安駟</a></li></ul>
                            </td>
                            <td align="right" class=table_eng_text style="width: 52px; padding: 4px 3px; vertical-align: middle;">(104)</td>
                        </tr>
                    </table>
                </td>
                <td width=20% bgcolor=FFFFFF class=table_eng_text>
                    <table width="100%" cellspacing="0" cellpadding="0" border=0>
                        <tr>
                            <td class=table_text_two bgcolor=#ffffff align=left style="padding:4px 2px; width:100px;">
                                <ul><li class=table_text_two style="padding:4px 2px"><a href="/racing/information/chinese/Horse/Horse.aspx?HorseId=HK_2022_H411" class=ltable_text>氣勢</a></li></ul>
                            </td>
                            <td align="right" class=table_eng_text style="width: 52px; padding: 4px 3px; vertical-align: middle;">(98)</td>
                        </tr>
                    </table>
                </td>
                <td width=20% bgcolor=FFFFFF class=table_eng_text>
```

Each horse

```html
<table class="horseProfile">
    <tbody>
        <tr>
            <td style="width: 280px;" valign="top">
                <table style="width: 280px; border:0;">
                    <tbody>
                        <tr>
                            <td class="subsubheader" colspan="2"><span class="title_text">安遇 (G236)</span></td>
                        </tr>
                        <tr>
                            <td style="width: 105px;" valign="top">
                                <table id="Table3" cellspacing="0" cellpadding="0">
                                    <tbody>
                                        <tr>
                                            <td><a onclick="window.open(this.href,'_blank','height=330,width=370');return false;" href="/racing/chinese/horse_img.asp?filename=G236_l.jpg"><img src="/racing/content/Images/horse/G236_s.jpg" onerror="this.style.opacity='0'; this.parentNode.removeAttribute('onclick'); this.parentNode.removeAttribute('href');
                                        </tr>
                                    </tbody>
                                </table>
                                <table id="Table4" class="icon_margin" width="105" cellspacing="0" cellpadding="0">
                                    <tbody>
                                        <tr>
                                            <td align="center" bgcolor="#EAEAEA" height="21"><img src="/racing/content/Images/RaceColor/G236.gif" onerror="this.style.opacity='0';" alt="" width="30" height="38" /></td>
                                        </tr>
                                    </tbody>
                                </table>
                                    <div class="PENote_HK_2021_G236"></div>
                            </td>
                            <td class="table_text_two" style="min-width: 145px; padding: 0 0 0 20px;">
                                <ul>
                                    <li><a class="table_text_two" href="/racing/information/Chinese/Horse/Horse.aspx?HorseId=HK_2021_G236">往績紀錄</a></li>
                                    <li><a class="table_text_two" href="/racing/information/Chinese/Horse/RatingResultWeight.aspx?HorseId=HK_2021_G236">馬匹評分/體重/名次</a></li>
                                    <li><a class="table_text_two" href="/racing/information/Chinese/Horse/Performance.aspx?HorseId=HK_2021_G236">所跑途程賽績紀錄</a></li>
                                    <li><a class="table_text_two" href="/racing/information/Chinese/Trackwork/TrackworkResult.aspx?HorseId=HK_2021_G236">晨操紀錄</a></li>
                                    <li><a class="table_text_two" href="/racing/information/Chinese/VeterinaryRecords/OveHorse.aspx?HorseId=HK_2021_G236">傷患紀錄</a></li>
                                        <li><a class="table_text_two" href="/racing/information/Chinese/Horse/MovementRecords.aspx?HorseId=HK_2021_G236">搬遷紀錄</a></li>
                                                                <li><a class="table_text_two" href="/racing/content/chinese/pp_formsheet/fsc_G236.pdf" target="_blank">海外賽績紀錄</a></li>
                                                                <li><a class="table_text_two" href="/racing/chinese/racing-info/newhorse.asp?racedate=20220518&amp;raceNo=6&amp;brandNo=G236">血統簡評</a></li>
                                    <li><a class="table_text_two" href="/racing/information/Chinese/Horse/SelectHorse.aspx">其他馬匹</a></li>
                                </ul>
                            </td>
                        </tr>
                    </tbody>
                </table>
            </td>
            <td valign="top" style="width: 260px;">
                <table style="width: 260px; border: 0;" class="table_top_right table_eng_text">
                    <tbody style="vertical-align: top; white-space:nowrap">
                        <tr>
                            <td>出生地 / 馬齡</td>
                            <td>:</td>
                            <td>愛爾蘭 / 5</td>
                        </tr>
                        <tr>
                            <td>毛色 / 性別</td>
                            <td>:</td>
                            <td>棕 / 閹</td>
                        </tr>
                        <tr>
                            <td>進口類別</td>
                            <td>:</td>
                            <td>自購馬</td>
                        </tr>
                        <tr>
                            <td>今季獎金*</td>
                            <td>:</td>
                            <td>$187,250</td>
                        </tr>
                        <tr>
                            <td>總獎金*</td>
                            <td>:</td>
                            <td>$14,181,025</td>
                        </tr>
                        <tr>
                            <td>冠-亞-季-總出賽次數*</td>
                            <td>:</td>
                                <td>7-2-2-28</td>
                        </tr>
                        <tr>
                            <td>最近十個賽馬日<br />出賽場數</td>
                            <td>:</td>
                            <td>2</td>
                        </tr>
                                                  <tr>
                                <td>現在位置<br />(到達日期)</td>
                                <td>:</td>
                                <td>香港<br />(24/09/2024)</td>
                            </tr>
                            <tr>
                                <td>進口日期</td>
                                <td>:</td>
                                <td>16/10/2021</td>
                            </tr>
```

Specific information

Racing records

Data scraping process.

(1) Webpage structure and its techniques

The Hong Kong Jockey Club categorized horse information by the number of characters in the horse names (such as "二字馬" and "三字馬"), providing a main URL for horse selection according to character count. This structure includes HTML tables that display the horse names and links to detailed pages for each horse, where relevant information is organized within specific HTML tables.

(2) Data Scraping Methods

During the data scraping process, we used the following Python libraries:

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import time
```

Requests: Used to send HTTP requests to retrieve the HTML content of a webpage.

BeautifulSoup: used to parse HTML and extract the required data.

Pandas: Used to organize the scraped data into a DataFrame and save it as a CSV file.

Time: To manage delays between requests and avoid overwhelming the server.

During the scraping process, we first define the categories of horses, build URLs for each type of horse, and send a GET request to visit the corresponding page. Here are the main steps in the scraping process:

Send request: Use requests.get() to send HTTP requests and obtain responses, and use headers to simulate browser requests to avoid being recognized by websites as crawlers.

```
# 发送 HTTP GET 请求以获取马匹列表
headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/
}
response = requests.get(trainer_url, headers=headers)
```

HTML parsing: Use BeautifulSoup to parse HTML and use the. find() and. find_all() methods to locate specific HTML elements.

```
# 解析 HTML 内容
soup = BeautifulSoup(response.text, 'html.parser')

# 提取所有马匹的信息
horse_links = []

# 查找当前马匹类型部分
horse_table = soup.find('td', class_='subheader', string=f'{horse_name} (依評分由大至小横向排列)')
if horse_table:
    # 找到包含具体马匹信息的表格
    inner_table = horse_table.find_next('table')
    rows = inner_table.find_all('tr')
```

Iteratively capture horse information: Iterate through each horse's links to get its detailed information.

```
# 遍历每匹马并提取详细信息
unique_horse_names = set()
for horse_name, horse_url in horse_links:
    if horse_name in unique_horse_names:
        continue  # 如果马匹名称已存在，跳过
    unique_horse_names.add(horse_name)  # 添加到集合中，以防重复
```

```
# 解析马匹基本信息
horse_info = {'马匹名称': horse_name}

details = horse_soup.find_all('table', class_='table_top_right table_eng_text')
for detail in details:
    rows = detail.find_all('tr')
    for row in rows:
        cols = row.find_all('td')
        if len(cols) >= 3:
            key = cols[0].get_text(strip=True)
            value = cols[2].get_text(strip=True)
            horse_info[key] = value
```

```
# 提取近三季往绩记录
performance_table = horse_soup.find('table', class_='bigborder')
if performance_table:
    records = performance_table.find_all('tr')[2:]  # 跳过前两行表头
```

Data storage: Store the scraped information in a list, and finally, use Pandas to save the data as a CSV file.

(3) Scraping results.

In the end, we captured a total of 1235 horses in Hong Kong, along with their recent 3 seasons records:

马匹名称	生地 / 馬	色 / 性	進口類別	今季獎金*	總獎金*	今季-總出賽	現在馬日	位置(到達	進口日期	練馬師	馬主	現時評分	季初評分	父系	母系	外祖父	同父系馬
安遇	愛爾蘭 / 棕 / 閹	自購馬	$0	$13,993,7	7-2-2-27	2		香港(24/0	16/10/202	曾誉	王賢訊	104	107	Churchill	Enrol	Pivotal	美麗之光
安騁	美國 / 4	栗 / 閹	自購馬	$3,975,00	$11,668,4	4-1-1-9	2		香港(21/0	10/06/202	蔡約翰	王賢敏	104	85	Lemon Dro	Alluvial	Danehill	沒有
氣勢	澳洲 / 4	栗 / 閹	自購新馬	$4,228,40	$8,012,60	6-2-0-8	2		香港(19/0	08/05/202	蔡約翰	李國能	98	80	Rich Enuf	Could Be	General N	跨境寶馬
增有	澳洲 / 6	栗 / 閹	自購馬	$0	$10,571,5	4-3-3-31	2		香港(07/1	13/09/202	曾誉	曾永達	91	95	Street Bo	Kellys Oc	Dansili	K174
球星	澳洲 / 4	栗 / 閹	自購新馬	$1,041,60	$4,819,20	4-2-5-14	1		香港(22/1	08/05/202	大衛希斯	香港足球	83	76	Rubick	Emirates	Dubawi	星河勇士
瑪瑙	澳洲 / 6	栗 / 閹	自購馬	$232,800	$9,468,80	3-3-1-26	1		從化(13/1	20/12/202	羅富全	Chong M	82	85	All Too H	Urban Roc	Fastnet R	精鵬燒
晴王	紐西蘭 / 栗	閹	自購新馬	$140,400	$6,294,15	5-1-3-21	1		香港(07/0	22/03/202	鄭俊偉	吳廷傑	78	78	Frosted	Darci Dan	Darci Bra	八勇士多
耀寶	澳洲 / 4	栗 / 閹	自購馬	$3,583,20	$3,583,20	2-0-0-3	2		香港(04/0	18/03/202	羅富全	許晉亨	77	64	Toronado	Liberty M	Statue Of	雄彬飛翔
玩笑	英國 / 3	栗 / 閹	自購馬	$0	$0	0-0-0-0	0		香港(24/1	06/07/202	伍鵬志	游國輝	76	76	Siyouni	Illaunmor	Shamardal	縱橫大進
米奇	愛爾蘭 / 栗	閹	自購馬	$0	$0	0-0-0-0	0		香港(04/0	04/08/202	蔡約翰	蕭劍平	76	76	Soldier's	Parle Moi	Giant's C	沒有
優才	澳洲 / 4	栗 / 閹	自購馬	$0	$0	0-0-0-1	1		香港(24/0	24/06/202	葉楚航	梁麟炳	73	73	Real Stee	Northern	Northern	幸運震撼
將義	紐西蘭 / 栗	閹	自購馬	$626,350	$698,100	0-1-1-4	1		香港(02/0	21/03/202	韋達	七俠五義	72	70	U S Navy	Pinstripe	Pins	時時如意
十力	愛爾蘭 / 栗	閹	自購馬	$0	$0	0-0-0-0	0		香港(05/1	05/10/202	伍鵬志	誠信團體	72	72	Justify	Hence	Galileo	風再起時
馬力	澳洲 / 4	栗 / 閹	自購馬	$2,541,60	$2,541,60	1-0-0-1	1		香港(18/0	18/03/202	羅富全	利子厚	71	65	Harry Ang	Romneya	Red Ranso	天使之聲
祝願	澳洲 / 4	栗 / 閹	自購新馬	$1,910,70	$2,197,35	2-1-1-5	1		香港(19/1	19/12/202	廖康銘	車曉紅	71	54	Flying Ar	Set The T	Reset	美麗老
豐盛	澳洲 / 4	栗 / 雄	自購馬	$653,700	$653,700	0-1-0-5	2		香港(15/0	01/03/202	廖康銘	多多利	71	69	National	Choice Wo	Choisir	星運遨翔
快路	愛爾蘭 / 栗	閹	自購馬	$374,125	$3,718,32	2-1-0-13	2		從化(11/1	05/11/202	告東尼	黃永光	70	69	Galileo G	Providenc	Acclamati	沒有
飲杯	澳洲 / 4	栗 / 閹	自購新馬	$1,324,50	$2,225,40	2-1-1-8	1		香港(26/1	07/11/202	大衛希斯	黃志祥	69	59	Fastnet R	My Sabeel	Savabeel	星儔領創
伊臣	澳洲 / 6	栗 / 閹	自購新馬	$213,900	$7,505,75	5-3-1-28	1		香港(26/0	14/07/202	姚本輝	李伊湄	69	69	Written T	Prete A P	Choisir	皆益善佳
增勁	英國 / 4	栗 / 閹	自購馬	$0	$0	0-0-0-1	1		香港(29/0	14/10/202	曾誉	曾永達	69	69	Lope de V	Suffused	Champs El	悠悠乾坤
俊才	愛爾蘭 / 栗	閹	自購馬	$0	$0	0-0-0-0	0		香港(06/0	06/07/202	蘇偉賢	梁麟炳	68	68	Phoenix O	Alfea	Kentucky	祥勝將軍
銀進	澳洲 / 6	栗 / 閹	自購新馬	$1,305,95	$5,527,80	4-4-4-31	1		香港(09/1	09/11/202	方嘉柏	陳永光	67	60	Hinchinbr	Morinda	Desert Ki	沒有
奔放	愛爾蘭 / 栗	閹	自購馬	$0	$0	0-0-0-0	0		從化(21/1	07/09/202	告東尼	蕭劍莹	67	67	Muhaarar	Bahjtee	Pivotal	沒有
堅閹	紐西蘭 / 棕	閹	自購新馬	$1,454,40	$2,554,20	2-2-2-13	1		香港(11/0	29/06/202	伍鵬志	張顯信	66	57	War Decre	Udiditaga	D' Cash	穿甲戰
增強	澳洲 / 4	栗 / 閹	自購新馬	$1,444,95	$2,205,45	2-2-3-7	1		從化(08/1	28/08/202	曾誉	曾永達	66	52	Capitalis	Sebrina	Sebring	喆喆友福
快搏	澳洲 / 7	栗 / 閹	自購馬	$65,100	$4,865,17	3-2-4-37	1		從化(25/1	10/05/202	曾誉	鍾建書	66	70	Written T	Best Yet	Econsul	皆益善
佳登	愛爾蘭 / 栗	閹	自購馬	$0	$0	0-0-0-0	0		從化(18/1	07/09/202	游達荣	晴朗的	66	66	Kodi Bear	Malilla	Red Clubs	沒有
福星	澳洲 / 5	栗 / 閹	自購新馬	$0	$1,769,10	2-0-2-9	0		從化(11/1	07/11/202	曾誉	鄭永安	66	66	Star Turn	Madame Bo	Publishin	美麗緣

马匹名称	场次	名次	日期	赛道	途程	场地状况	赛事班次	档位	评分	练马师	骑师	头马距离	独赢赔率	实际负磅	沿途走位	完成时间	排位体重	配备
安遇	153	08	03/11/24	沙田草地	1800	好/快	G3	8	105	智誉	霍宏聲	3-1/2	42	121	12 12 12	1.46.97	1161	B
安遇	117	12	20/10/24	沙田草地	1200	好/快	G2	6	107	智誉	班德禮	7-1/2	183	116	12 13 12	1.08.79	1168	B2/TT-
安遇	719	07	02/06/24	沙田草地	1600	好/快	G3	4	111	文家良	田泰安	13-3/4	26	127	4 3 6 7	1.36.15	1115	TT
安遇	566	13	07/04/24	沙田草地	1600	好	G2	4	113	文家良	田泰安	17-1/4	10	123	7 6 10 1	1.36.79	1125	TT
安遇	452	07	25/02/24	沙田草地	2000	好	G1	2	114	文家良	田泰安	5-1/2	16	126	5 6 7 7	2.01.20	1146	TT
安遇	395	08	04/02/24	沙田草地	1800	好	G3	12	114	文家良	梁家俊	3	15	132	2 3 5 5	1.47.89	1155	TT
安遇	330	08	10/01/24	跑馬地草	1800	好	G3	12	114	文家良	梁家俊	4-1/4	11	135	7 8 8 7	1.48.99	1164	TT
安遇	238	08	10/12/23	沙田草地	1600	好	G1	1	114	文家良	梁家俊	4-3/4	24	126	6 6 7 8	1.34.88	1164	TT
安遇	144	01	05/11/23	沙田草地	1800	好	G3	6	110	文家良	梁家俊	1-1/4	12	128	3 4 4 4	1.47.40	1164	TT
安遇	088	02	15/10/23	沙田草地	1600	好	G2	4	105	文家良	梁家俊	1/2	14	117	3 3 4 2	1.34.28	1156	TT
安遇	039	06	24/09/23	沙田草地	1400	好	G3	6	105	文家良	梁家俊	10-1/2	4.4	123	5 6 4 6	1.23.59	1153	TT
安遇	834	01	16/07/23	沙田草地	1600	好	1	7	100	文家良	梁家俊	頸位	4.9	128	9 9 5 1	1.34.93	1165	TT
安遇	749	01	14/06/23	跑馬地草	1650	好	2	10	89	文家良	梁家俊	4-3/4	3.2	124	5 4 1 1	1.38.87	1153	TT
安遇	576	01	12/04/23	跑馬地草	1650	好	2	6	83	文家良	梁家俊	1	5.9	120	4 4 3 1	1.39.08	1138	TT
安遇	509	09	19/03/23	沙田草地	2000	好	4YO	8	83	文家良	蕭兆輝	3-3/4	124	126	2 2 2 2	2.03.37	1152	B-/H-/TT
安遇	453	12	26/02/23	沙田草地	1800	好/快	4YO	9	83	文家良	蕭兆輝	8-1/2	28	126	7 6 6 11	1.47.62	1152	B/H1/TT
安遇	406	02	08/02/23	跑馬地草	1650	好	3	5	81	文家良	蕭兆輝	頸位	2.2	132	2 2 2 2	1.39.62	1153	B/TT
安遇	328	01	11/01/23	跑馬地草	1650	好	3	12	74	文家良	蕭兆輝	1	4.2	131	2 2 1 1	1.39.53	1137	B/TT
安遇	250	01	14/12/22	跑馬地草	1650	好	3	3	67	文家良	蕭兆輝	2-1/4	3.5	126	2 3 4 1	1.39.13	1129	B1/TT
安遇	166	03	12/11/22	沙田全天	1650	好	3	6	68	文家良	潘明輝	9	11	121	2 2 2 3	1.38.22	1133	TT
安遇	116	05	23/10/22	沙田草地	1600	好/快	3	11	69	文家良	潘頓	2	5.8	124	2 2 2 5	1.33.80	1136	TT
安遇	044	06	25/09/22	沙田草地	1600	好/快	3	2	69	文家良	潘明輝	2-3/4	4.2	124	3 5 5 6	1.35.23	1143	TT
安遇	010	03	11/09/22	沙田草地	1400	好	3	12	69	文家良	潘明輝	1-1/4	3.4	122	4 2 2 3	1.22.58	1140	TT
安賽	153	01	03/11/24	沙田草地	1800	好/快	G3	13	86	蔡約翰	田泰安	3/4	11	115	3 4 5 5	1.46.41	1081	--
安賽	095	04	13/10/24	沙田草地	2000	好/快	3	5	85	蔡約翰	潘頓	1/2	2.3	135	3 4 4 3	2.01.52	1072	--
安賽	528	12	24/03/24	沙田草地	2000	好/快	4YO	12	85	蔡約翰	莫雅	16-1/2	6.8	126	5 5 5 11	2.02.49	1053	--
安賽	471	03	03/03/24	沙田草地	1800	好	4YO	12	85	蔡約翰	莫雅	2-3/4	7.9	126	4 3 2 4	1.48.17	1069	--
安賽	358	02	21/01/24	沙田草地	2000	好	2	4	83	蔡約翰	田泰安	頸位	2.9	131	3 4 3 3	2.02.60	1062	--
安賽	275	01	23/12/23	沙田草地	2000	好	2	9	76	蔡約翰	田泰安	3/4	2.7	115	5 4 4 3	2.01.77	1067	--
安賽	237	08	10/12/23	沙田草地	1800	好	2	2	76	蔡約翰	艾道拿	6-3/4	2.7	133	5 5 8 9	1.49.40	1047	--
安賽	181	01	19/11/23	沙田草地	2000	好/快	3	12	70	蔡約翰	艾道拿	頸位	4.8	126	6 6 5 6	2.00.93	1065	--
安賽	087	01	15/10/23	沙田草地	1800	好	3	11	64	蔡約翰	潘頓	1/2	7.8	123	6 6 5 2	1.48.03	1063	--
氣勢	171	02	09/11/24	沙田草地	1000	好	2	2	95	蔡約翰	潘頓	短馬頭位	1.4	132	2 2 2	0.56.39	1211	--
氣勢	102	01	13/10/24	沙田草地	1200	好/快	2	1	88	蔡約翰	潘頓	頸位	1.9	124	2 2 1	1.08.29	1218	--
氣勢	045	01	22/09/24	沙田草地	1200	黏	3	2	80	蔡約翰	潘頓	1/2	1.5	134	1 1 1	1.09.65	1202	--
氣勢	353	01	21/01/24	沙田草地	1200	好	3	2	73	蔡約翰	麥道朗	短馬頭位	2	129	2 2 1	1.08.93	1194	--
氣勢	305	01	01/01/24	沙田草地	1200	好	3	7	66	蔡約翰	麥道朗	鼻位	5.1	122	3 3 1	1.09.12	1201	--
氣勢	260	02	17/12/23	沙田草地	1200	好	3	9	65	蔡約翰	麥道朗	2-1/4	5.9	123	2 1 2	1.09.60	1195	--
氣勢	196	01	26/11/23	沙田草地	1000	好	4	4	58	蔡約翰	麥道朗	3/4	1.6	133	2 1 1	0.57.39	1199	--
氣勢	140	01	05/11/23	沙田草地	1000	好	4	5	52	蔡約翰	田泰安	頸位	10	123	4 2 1	0.56.29	1195	--
增有	172	09	09/11/24	沙田草地	1400	好/快	2	8	93	智誉	潘頓	3-3/4	24	123	12 11 10	1.20.93	1070	B/H/XB
增有	102	09	13/10/24	沙田草地	1200	好/快	2	2	95	智誉	布文	5-3/4	33	131	12 12 9	1.09.21	1068	B/H/XB
增有	828	04	14/07/24	沙田草地	1600	好	1	6	96	智誉	田泰安	4-3/4	8.8	126	10 10 9	1.34.55	1068	B/H/XB
增有	787	05	01/07/24	沙田草地	1400	好	2	4	97	智誉	布文	3-1/4	8.7	135	7 6 8 5	1.21.23	1064	B/H/XB
增有	719	05	02/06/24	沙田草地	1600	好/快	G3	2	97	智誉	巴度	3	8.2	115	3 4 3 5	1.34.43	1062	B/H/XB
增有	677	02	19/05/24	沙田草地	1400	好	1	6	97	智誉	潘頓	短馬頭位	3.6	122	6 6 6 2	1.21.59	1061	B/H/XB
增有	625	04	28/04/24	沙田草地	1400	黏	2	10	97	智誉	布文	1-1/4	7.6	132	13 12 11	1.22.55	1058	B/H/XB

W2-MaJing

More comprehensive race data on horse racing is available on the Majing website, which is the reason for choosing to crawl the data from the second website.

Targeted web pages are as follows:

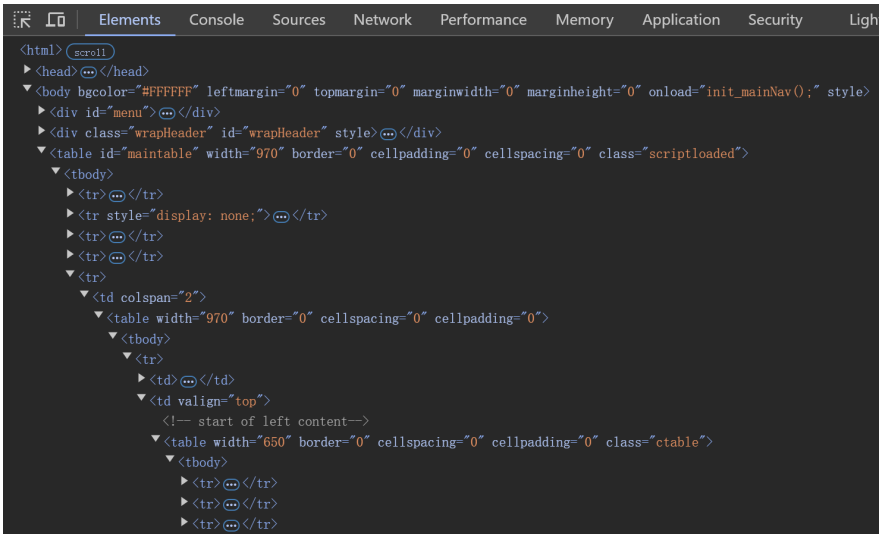
Detailed racing data

URL: https://racing.on.cc/cgi-bin/srh/search/search.cgi / (First website)

https://racing.on.cc/cgi-bin/srh/search/onlinesrh.cgi (Final website)

Web Page Structure of Majing website: HTML

The inspecting the element windows are as follows:



Data Scraping Process.

(1) Page Loading and Element Location

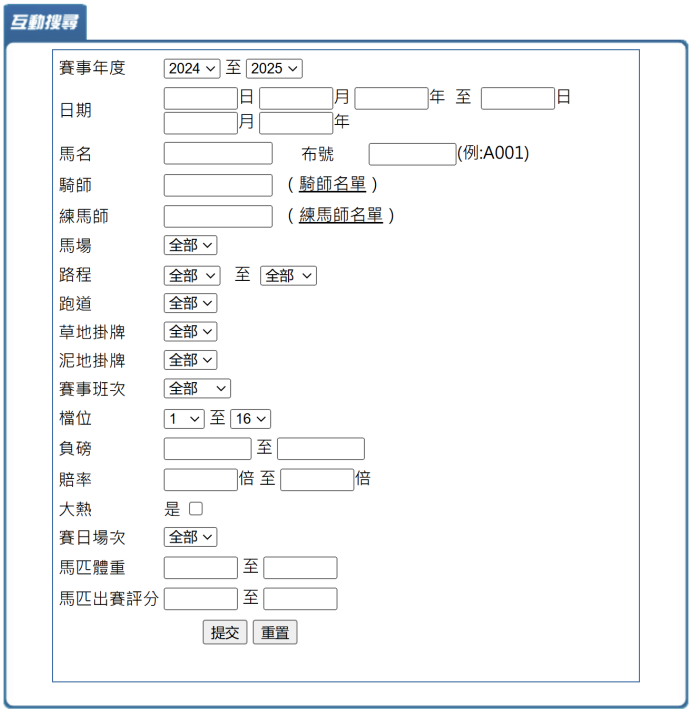
Since it is impossible to directly obtain the content of the target web page, we use the Selenium to crawl the web page data.

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
```

(2) Form Submission

Through the Elements, find the name of the start year, the end year, and the submit button

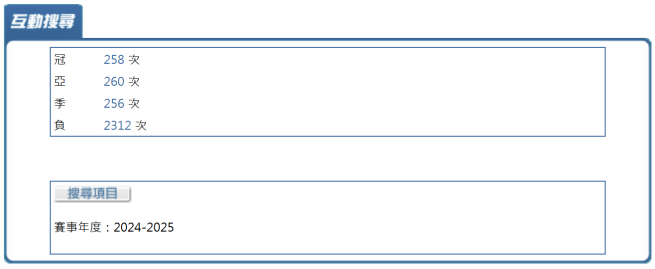
on the entry website. The script selects start and end years from dropdowns and submits the form by locating and clicking the submit button.



Accessed pages

(3) Navigation to Result Page

After submitting, it will be redirected to a second category page as follows:



After selecting and clicking on the specific links, it will be redirected to the target page.

(4) Data Extraction

The script locates a target table on the target page with Xpath and extracts rows and cells. Then, store the data in a list, and convert it into a Pandas DataFrame.

```
# 获取表格内容
try:
    # 使用提供的XPath定位表格
    table = driver.find_element(By.XPATH, '//*[@id="maintable"]/tbody/tr[5]/td/table/tbody/tr/td[2]/table/tbody/tr[4]/td[2]/table/tbody/tr/td/table/tbc
    rows = table.find_elements(By.TAG_NAME, "tr")  # 获取所有行
```

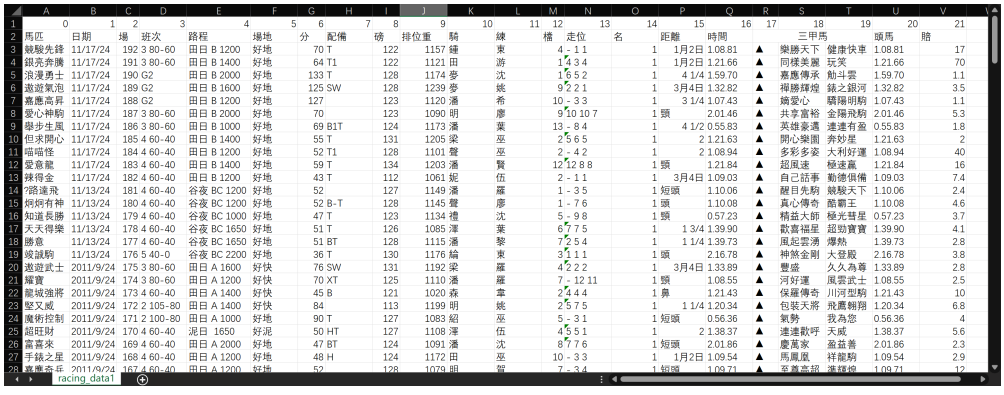
(5) Data Storage

The DataFrame is saved as a CSV file, ensuring proper encoding for international characters.

```
# 检查是否获取到数据
if table_data:
    # 将数据转换为DataFrame
    df = pd.DataFrame(table_data)

    # 保存为CSV文件，使用 utf-8-sig 编码，确保中文字符在Excel中正常显示
    df.to_csv("racing_data2.csv", index=False, encoding="utf-8-sig")

    print("数据已保存为racing_data.csv")
```

(6) Scraping results

Through the above process, we have collected nearly 30,000 pieces of race data from 4 seasons (2021-2024) of Hong Kong horse racing.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	馬匹	日期	場	班次	路程	場地	分	配備	磅	排位重	騎	練	檔	走位	名	距離	時間		三甲馬		頭馬	賠
2	競駿先鋒	11/17/24	192	3 80-60	田B 1200	好地	70	T	122	1157	鍾	東	4	- 1 1	1	1月2日	1.08.81	▲	樂勝天下	健康快車	1.08.81	17
3	銀亮奔騰	11/17/24	191	3 80-60	田B 1400	好地	64	T1	122	1121	田	游	1	4 3 4	1	1月2日	1.21.66	▲	同樑美麗	玩笑	1.21.66	70
4	浪漫勇士	11/17/24	190	G2	田B 2000	好地	133	T	128	1174	麥	沈	1	5 5 2	1	4 1/4	1.59.70	▲	嘉應傳承	勉斗雲	1.59.70	1.1
5	遨遊氣泡	11/17/24	189	G2	田B 1600	好地	125	SW	128	1239	麥	姚	9	2 2 1	1	3月4日	1.32.82	▲	禪勝輝煌	錶之銀河	1.32.82	3.5
6	嘉應高昇	11/17/24	188	G2	田B 1200	好地	127		123	1120	潘	希	10	- 3 3	1	3 1/4	1.07.43	▲	嫣愛心	騎陽明駒	1.07.43	1.1
7	愛心神駒	11/17/24	187	3 80-60	田B 2000	好地	70		123	1090	明	廖	9	10 10 7	1	頸	2.01.46	▲	共享富裕	金陽飛駒	2.01.46	5.3
8	舉步生風	11/17/24	186	3 80-60	田B 1000	好地	69	B1T	124	1173	潘	葉	13	- 8 4	1	4 1/2	0.55.83	▲	英雄豪邁	連連有盈	0.55.83	1.8
9	但求開心	11/17/24	185	4 60-40	田B 1400	好地	55	T	131	1205	梁	巫	2	5 6 5	1	2	1.21.63	▲	開心樂園	奔妙星	1.21.63	2
10	喵喵怪	11/17/24	184	4 60-40	田B 1200	好地	52	T1	128	1101	聲	巫	2	- 4 2	1	2	1.08.94	▲	多彩多姿	大利好運	1.08.94	40
11	愛意龍	11/17/24	183	4 60-40	田B 1400	好地	59	T	134	1203	潘	賢	12	12 8 8	1	頸	1.21.84	▲	超風速	極速贏	1.21.84	16
12	辣得金	11/17/24	182	4 60-40	田B 1200	好地	43	T	112	1061	妮	伍	2	- 1 1	1	3月4日	1.09.03	▲	自己話事	勤德俱備	1.09.03	7.4
13	飛路遠飛	11/13/24	181	4 60-40	谷夜 BC 1200	好地	52		127	1149	潘	羅	1	- 3 5	1	短頭	1.10.06	▲	醒目先駒	競駿天下	1.10.06	2.4
14	炳炳有神	11/13/24	180	4 60-40	谷夜 BC 1200	好地	52	B-T	128	1145	聲	廖	1	- 7 6	1	頭	1.10.08	▲	真心傳奇	酷霸王	1.10.08	4.6
15	知道長勝	11/13/24	179	4 60-40	谷夜 BC 1000	好地	47	T	123	1134	禮	沈	5	- 9 8	1	頸	0.57.23	▲	精益大師	極光彗星	0.57.23	3.7
16	天天得樂	11/13/24	178	4 60-40	谷夜 BC 1650	好地	51	T	126	1085	潭	葉	6	7 7 5	1	1 3/4	1.39.90	▲	歡喜福星	超勁寶寶	1.39.90	4.1
17	勝意	11/13/24	177	4 60-40	谷夜 BC 1650	好地	51	BT	128	1115	潘	黎	7	2 5 4	1	1 1/4	1.39.73	▲	風起雲湧	爆熱	1.39.73	2.8
18	競誠駒	11/13/24	176	5 40-0	谷夜 BC 2200	好地	36	T	130	1176	綸	東	3	1 1 1	1	頭	2.16.78	▲	神煞金剛	大登殿	2.16.78	3.8
19	遨遊武士	2011/9/24	175	3 80-60	田B A 1600	好快	76	SW	131	1192	梁	羅	4	2 2 2	1	3月4日	1.33.89	▲	豐盛	久久為尊	1.33.89	2.8
20	耀寶	2011/9/24	174	3 80-60	田B A 1200	好快	70	XT	125	1110	潘	羅	7	- 12 11	1	頸	1.08.55	▲	河好運	風雲武士	1.08.55	2.5
21	龍城強將	2011/9/24	173	4 60-40	田B A 1400	好快	45	B	121	1020	森	韋	2	4 4 4	1	鼻	1.21.43	▲	保羅傳奇	川河型駒	1.21.43	10
22	堅又威	2011/9/24	172	2 105-80	田B A 1400	好快	84		113	1199	明	姚	2	5 7 5	1	1 1/4	1.20.34	▲	包裝天將	飛鷹翱翔	1.20.34	6.8
23	魔術控制	2011/9/24	171	2 100-80	田B A 1000	好地	90	T	127	1083	紹	巫	5	- 3 1	1	短頭	0.56.36	▲	氣勢	我為您	0.56.36	4
24	超旺財	2011/9/24	170	4 60-40	泥B 1650	好泥	50	HT	127	1108	澤	伍	4	5 5 1	1	2	1.38.37	▲	連連歡呼	天威	1.38.37	5.6
25	富喜來	2011/9/24	169	4 60-40	田B A 2000	好地	47	BT	124	1091	潘	沈	8	7 7 6	1	短頭	2.01.86	▲	慶萬家	盈益善	2.01.86	2.3
26	手錶之星	2011/9/24	168	4 60-40	田B A 1200	好地	48	H	124	1172	田	巫	10	- 3 3	1	1月2日	1.09.54	▲	馬鳳凰	祥龍駒	1.09.54	2.9
27	嘉應奇兵	2011/9/24	167	4 60-40	田B A 1200	好地	52		128	1079	明	賀	7	- 3 4	1	短頭	1.09.71	▲	至尊高昇	蒲點將	1.09.71	12

racing_data1

The raw dataset

Difficulties and Solutions

Although the Majing website has no obvious anti-crawl mechanism, we still encountered a few minor problems in crawling the data.

(1) Simulator jumping.

Difficulty: During the crawling process, we can't reach the target page directly because copying the URL of the target page directly will only result in a blank page without data. Therefore, in order to ensure that the data is visible, we can only access the target page by searching for the jumps.

Pages accessed by copying URL

Solution: Use selenium to simulate a human search from the search portal and reach our target page through two webpage jumps to get our target data.

```python
# Step 1: 打开搜索页面
driver.get("https://racing.on.cc/cgi-bin/srh/search/search.cgi")
time.sleep(2)  # 等待页面加载

# Step 2: 找到“赛事年度（开始年份）”选择框并选择"2024"
start_year_select = WebDriverWait(driver, 10).until(
    EC.presence_of_element_located((By.NAME, "fryear"))  # 根据实际的字段名称来获取
)
start_year_select = Select(start_year_select)
start_year_select.select_by_visible_text("2024")  # 选择2024

# Step 3: 找到“赛事年度（结束年份）”选择框并选择"2025"
end_year_select = WebDriverWait(driver, 10).until(
    EC.presence_of_element_located((By.NAME, "toyear"))  # 根据实际的字段名称来获取
)
end_year_select = Select(end_year_select)
end_year_select.select_by_visible_text("2025")  # 选择2025

# Step 4: 找到提交按钮，先使用滚动和显式等待确保按钮可点击
submit_button = WebDriverWait(driver, 10).until(
    EC.element_to_be_clickable((By.NAME, "submitButton"))
)
```

```python
# Step 5: 等待页面跳转到结果页面
time.sleep(3)  # 等待页面跳转

# Step 6: 在结果页面中找到链接文本为"192"的元素并点击
try:
    # 根据链接文本来找到并点击
    link = WebDriverWait(driver, 10).until(
        EC.presence_of_element_located((By.LINK_TEXT, "1692"))  # 查找链接文本为"192"的元素
    )
    link.click()  # 点击链接

    # 等待页面跳转
    time.sleep(3)  # 等待页面加载到目标页面
```

(2) Ad blocking.

Difficulty: In the search page, affected by pop-up ads, unable to successfully click to the submit button.

Solution: Use code to make sure the button is visible.

```python
# Step 4: 找到提交按钮，先使用滚动和显式等待确保按钮可点击
submit_button = WebDriverWait(driver, 10).until(
    EC.element_to_be_clickable((By.NAME, "submitButton"))
)

# 确保按钮可见
driver.execute_script("arguments[0].scrollIntoView();", submit_button)
```

Data Processing

MaJing

(1) Integrate four datasets from the website.

Because we have obtained the data of "冠", "亚", "季" and "亚" from MaJing, we need to integrate the data.

```python
import pandas as pd

# File paths
file_paths = [
    '/Users/Keanu/Desktop/racing_data/csv/冠.csv',
    '/Users/Keanu/Desktop/racing_data/csv/亚.csv',
    '/Users/Keanu/Desktop/racing_data/csv/季.csv',
    '/Users/Keanu/Desktop/racing_data/csv/负.csv'
]

# Load each CSV file into a DataFrame and concatenate them
combined_all_df = pd.concat([pd.read_csv(file) for file in file_paths], axis=0, ignore_index=True).drop_duplicates()

# Set output file path and export the combined DataFrame to CSV
output_file_path = '/Users/Keanu/Desktop/all.csv'
combined_all_df.to_csv(output_file_path, index=False, encoding='utf-8')

print(f"The combined table has been saved to: '{output_file_path}'")
```

(2) Sort by Date(日期) & Event(场) In Descending Order.

The original dates were mixed up and needed to be reorganized.

```python
import pandas as pd

# Load the CSV file
file_path = '/Users/Keanu/Desktop/all.csv'
data = pd.read_csv(file_path)

# Convert '日期' column to datetime
data['日期'] = pd.to_datetime(data['日期'], format='%m/%d/%y', errors='coerce')

# Drop rows with invalid dates
data = data.dropna(subset=['日期'])

# Sort by '日期' in descending order
df_sorted = data.sort_values(by='日期', ascending=False)

# Sort by '場' in descending order while preserving the order of '日期'
df_sorted = df_sorted.sort_values(by=['日期', '場'], ascending=[False, False])

# Reset the index
df_sorted.reset_index(drop=True, inplace=True)

# Save the sorted data to a new CSV file
df_sorted.to_csv('/Users/Keanu/Desktop/all_sorted.csv', index=False)

print("Data has been sorted and saved as 'all_sorted.csv'")
```

(3) Check Missing Values.

```python
import pandas as pd

# Load the CSV file
file_path = '/Users/Keanu/Desktop/all_sorted.csv'
data = pd.read_csv(file_path)

# Check for missing values in each column
missing_values = data.isnull().sum()
print("Missing values in each column:\n", missing_values)

# Extract rows with missing values
rows_with_missing = data[data.isnull().any(axis=1)]

# Save the rows with missing values to a new CSV file
rows_with_missing.to_csv('/Users/Keanu/Desktop/rows_with_missing_values.csv', index=False)

print("Rows with missing values have been saved as 'rows_with_missing_values.csv'")
```

(4) Follow the Result.

```
Missing values in each column:
 馬匹        0
日期         0
場          0
班次         0
路程         0
場地        10
分          0
配備      5061
磅          0
排位重        0
騎          0
練          0
檔          0
走位         0
名          0
距離        21
時間        25
冠          0
亞          0
季          0
頭馬        12
賠          0
dtype: int64
Rows with missing values have been saved as 'rows_with_missing_values.csv'
```

(5) Replace missing values in the “配备” column with None. According to the material and actual competitions, it is normal to have no equipment, so the "equipment" column is processed as "None", and then drop rows with other missing values.

```python
import pandas as pd

# Load the CSV file
file_path = '/Users/Keanu/Desktop/all_sorted.csv'
data = pd.read_csv(file_path)

# Replace missing values in '配備' column with 'None'
data['配備'].fillna('None', inplace=True)

# Drop rows with any other missing values
data = data.dropna()

# Remove rows where '名' column is equal to 90
data = data[data['名'] != 90]

# Check for missing values in each column
missing_values = data.isnull().sum()
print("Missing values in each column:\n", missing_values)

# Save the cleaned data to a new CSV file
data.to_csv('/Users/Keanu/Desktop/all_sorted_cleaned.csv', index=False)

print("Cleaned data has been saved as 'all_sorted_cleaned.csv'")
```

(6) Remove all other rows with missing values.

(7) Output the cleaned data.

cleaned_all_sorted

表格数据已导入且可调整。

馬匹	日期	場	班次	路程	場地	分	配備	磅	排位重	騎	練	檔	走位	名	距離	時間	冠	亞	季	頭馬	賠
得道猴王	2024-11-20	201	3 80-60	谷夜 C3 1200	好點	69	T	126	1196	鍾	東	10	- 1 1	1	頭	1.10.17	▲	伊臣	勇威神駒	1.10.17	4.4
寶貿得得	2024-11-20	201	3 80-60	谷夜 C3 1200	好點	75	T	128	1129	弘	柏	11	- 12 12	11	4 1/2	1.10.88	得道猴王	伊臣	勇威神駒	1.10.17	136.0
金發盛世	2024-11-20	201	3 80-60	谷夜 C3 1200	好點	66	B	126	1067	潭	呂	8	- 6 8	10	3 1/2	1.10.72	得道猴王	伊臣	勇威神駒	1.10.17	36.0
多利神駒	2024-11-20	201	3 80-60	谷夜 C3 1200	好點	62	T	122	1034	禮	賢	7	- 11 11	4	1 1/4	1.10.38	得道猴王	伊臣	勇威神駒	1.10.17	35.0
金運來	2024-11-20	201	3 80-60	谷夜 C3 1200	好點	60	XT	120	1102	森	文	12	- 2 2	6	2	1.10.47	得道猴王	伊臣	勇威神駒	1.10.17	28.0
無限美麗	2024-11-20	201	3 80-60	谷夜 C3 1200	好點	65	None	125	1153	布	蔡	4	- 9 9	5	1 3/4	1.10.43	得道猴王	伊臣	勇威神駒	1.10.17	8.6
多利彩駒	2024-11-20	201	3 80-60	谷夜 C3 1200	好點	69	XT	129	1085	班	廖	6	- 4 5	9	3 1/2	1.10.72	得道猴王	伊臣	勇威神駒	1.10.17	6.3
金牌活力	2024-11-20	201	3 80-60	谷夜 C3 1200	好點	63	B	123	1147	湯	鄭	5	- 10 10	7	2 3/4	1.10.61	得道猴王	伊臣	勇威神駒	1.10.17	63.0
友愛心得	2024-11-20	201	3 80-60	谷夜 C3 1200	好點	66	SW	126	1185	湛	希	9	- 3 3	12	7 1/4	1.11.33	得道猴王	伊臣	勇威神駒	1.10.17	7.7
競駿皇者	2024-11-20	201	3 80-60	谷夜 C3 1200	好點	74	T	134	1075	聲	游	3	- 5 4	8	3 1/4	1.10.70	得道猴王	伊臣	勇威神駒	1.10.17	5.5
伊臣	2024-11-20	201	3 80-60	谷夜 C3 1200	好點	69	None	127	1145	明	姚	2	- 7 6	2	頭	1.10.20	得道猴王	▲	勇威神駒	1.10.17	9.4
勇威神駒	2024-11-20	201	3 80-60	谷夜 C3 1200	好點	66	T	126	1137	麥	沈	1	- 8 7	3	1	1.10.34	得道猴王	伊臣	▲	1.10.17	5.5
美麗奔馳	2024-11-20	200	2 100-80	谷夜 C3 1200	好點	90	T	126	1096	艾	東	7	- 10 9	1	頸	1.09.43	▲	傑出漢子	精算暴雪	1.09.43	26.0
傑出漢子	2024-11-20	200	2 100-80	谷夜 C3 1200	好點	83	B	119	1097	森	呂	1	- 6 3	2	頸	1.09.47	美麗奔馳	▲	精算暴雪	1.09.43	20.0
我為您	2024-11-20	200	2 100-80	谷夜 C3 1200	好點	80	None	114	1166	明	伍	6	- 1 1	7	3 3/4	1.10.04	美麗奔馳	傑出漢子	精算暴雪	1.09.43	29.0
輝煌精英	2024-11-20	200	2 100-80	谷夜 C3 1200	好點	83	H	119	1220	梁	丁	10	- 2 2	8	4	1.10.06	美麗奔馳	傑出漢子	精算暴雪	1.09.43	9.3
占士德	2024-11-20	200	2 100-80	谷夜 C3 1200	好點	84	T1SW	120	1111	周	容	12	- 4 4	11	7	1.10.56	美麗奔馳	傑出漢子	精算暴雪	1.09.43	64.0
華麗再勝	2024-11-20	200	2 100-80	谷夜 C3 1200	好點	88	T	124	1160	麥	沈	8	- 7 7	5	2 1/4	1.09.80	美麗奔馳	傑出漢子	精算暴雪	1.09.43	9.4
昇灘駒	2024-11-20	200	2 100-80	谷夜 C3 1200	好點	82	None	118	1188	班	廖	3	- 8 6	4	2 1/4	1.09.78	美麗奔馳	傑出漢子	精算暴雪	1.09.43	7.4
精算暴雪	2024-11-20	200	2 100-80	谷夜 C3 1200	好點	88	P	124	1091	湛	蔡	4	- 11 10	3	1 1/4	1.09.64	美麗奔馳	傑出漢子	▲	1.09.43	2.2
幸運遇見	2024-11-20	200	2 100-80	谷夜 C3 1200	好點	91	None	127	1122	禮	希	2	- 5 8	12	7 3/4	1.10.66	美麗奔馳	傑出漢子	精算暴雪	1.09.43	5.9
小霸王	2024-11-20	200	2 100-80	谷夜 C3 1200	好點	99	BT	132	1188	鍾	東	9	- 12 12	6	3 3/4	1.10.04	美麗奔馳	傑出漢子	精算暴雪	1.09.43	15.0
魅力寶駒	2024-11-20	200	2 100-80	谷夜 C3 1200	好點	92	None	128	1197	度	蔡	5	- 9 11	9	4 1/2	1.10.16	美麗奔馳	傑出漢子	精算暴雪	1.09.43	28.0
人和家興	2024-11-20	200	2 100-80	谷夜 C3 1200	好點	94	BT	130	1202	董	希	11	- 3 5	10	5 1/2	1.10.31	美麗奔馳	傑出漢子	精算暴雪	1.09.43	45.0
勁好運	2024-11-20	199	3 80-60	谷夜 C3 1650	好點	62	T	120	1094	湛	羅	6	6 5 4	2	短頭	1.39.60	開心勇駒	▲	浪漫老搵	1.39.62	2.8
浪漫老搵	2024-11-20	199	3 80-60	谷夜 C3 1650	好點	72	SW2T	129	1152	梁	巫	2	8 8 8	3	1 1/4	1.39.80	開心勇駒	勁好運	▲	1.39.62	16.0
好好心得	2024-11-20	199	3 80-60	谷夜 C3 1650	好點	74	T-B	131	1080	禮	游	4	9 9 11	8	4 3/4	1.40.38	開心勇駒	勁好運	浪漫老搵	1.39.62	46.0
飛輪步	2024-11-20	199	3 80-60	谷夜 C3 1650	好點	74	T	131	1159	布	容	10	4 4 5	11	12 1/4	1.41.55	開心勇駒	勁好運	浪漫老搵	1.39.62	47.0
獨步天下	2024-11-20	199	3 80-60	谷夜 C3 1650	好點	63	None	120	1137	潭	呂	8	7 6 6	10	7 1/2	1.40.79	開心勇駒	勁好運	浪漫老搵	1.39.62	49.0

Hong Kong Jockey Club

(1) Data format conversion. Separate the mixed up information about total appearances and ranking information.

```python
import pandas as pd

# 读取csv文件
file_path = '/Users/Keanu/Desktop/racing_data/csv/HK_Horses.csv'
df = pd.read_csv(file_path)

# 定义一个函数来处理"冠-亞-季-總出賽次數+"列
def process_race_counts(race_counts_str):
    if pd.isna(race_counts_str):  # 处理缺失值
        return 0, 0, 0, 0  # 返回默认值

    try:
        parts = race_counts_str.split('-')
        if len(parts) == 4:
            wins = int(parts[0])
            seconds = int(parts[1])
            thirds = int(parts[2])
            total_races = int(parts[3].replace("+", "")) # Remove '+' and convert to int
            return wins, seconds, thirds, total_races
        else:
            return 0, 0, 0, 0 # handle cases with incorrect format
    except ValueError: # handle cases with non-numeric data
        return 0, 0, 0, 0


# 应用函数处理"冠-亞-季-總出賽次數+"列，并创建新列
df[['wins', 'seconds', 'thirds', 'total_races']] = df['冠-亞-季-總出賽次數+'].apply(lambda x: pd.Series(process_race_counts(x)))

# 删除原始的 '冠-亞-季-總出賽次數+' 列
df = df.drop('冠-亞-季-總出賽次數+', axis=1)

# 保存新的DataFrame到csv文件
new_file_path = '/Users/Keanu/Desktop/processed_data.csv'
df.to_csv(new_file_path, index=False)

print(f"Processed data saved to {new_file_path}")
```

(2) Check missing values.

```python
import pandas as pd

# Load the csv file
file_path = '/Users/Keanu/Desktop/processed_data.csv'
data = pd.read_csv(file_path)

# Check for missing values in each column
missing_values = data.isnull().sum()
print("Missing values in each column:\n", missing_values)

# Extract rows with missing values
rows_with_missing = data[data.isnull().any(axis=1)]

# Save the rows with missing values to a new CSV file
rows_with_missing.to_csv('/Users/Keanu/Desktop/rows_with_missing_values.csv', index=False)

print("Rows with missing values have been saved as 'rows_with_missing_values.csv'")
```

(3) Follow the result.

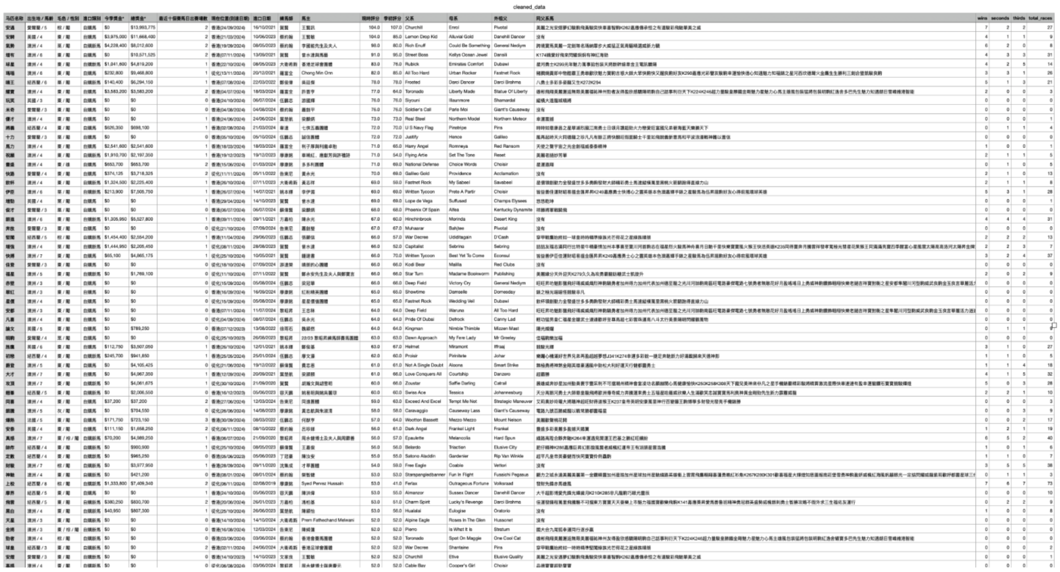
```
Missing values in each column:
 马匹名称                0
出生地 / 馬齡            0
毛色 / 性別              0
進口類別                 0
今季獎金*                0
總獎金*                  0
最近十個賽馬日出賽場數        0
現在位置(到達日期)           0
進口日期                 0
練馬師                  0
馬主                   0
現時評分                15
季初評分                15
父系                   0
母系                   0
外祖父                  0
同父系馬                 0
自購馬來港前賽事片段        1235
wins                 0
seconds              0
thirds               0
total_races          0
dtype: int64
Rows with missing values have been saved as 'rows_with_missing_values.csv'
```

(4) Remove all rows with missing values.

Because the column "自購馬來港前賽事片段" appears as a video link in the original webpage, it needs to be deleted first. Then drop rows with any other missing values.

```python
import pandas as pd

# Load the csv file
file_path = '/Users/Keanu/Desktop/processed_data.csv'
data = pd.read_csv(file_path)

# Remove the specified column
data = data.drop(columns=['自購馬來港前賽事片段'])

# Drop rows with any other missing values
data = data.dropna()

# Save the cleaned data to a new CSV file
output_file_path = '/Users/Keanu/Desktop/cleaned_data.csv'
data.to_csv(output_file_path, index=False)

# Check for missing values in each column
missing_values = data.isnull().sum()
print("Missing values in each column:\n", missing_values)
```

(5) Output the cleaned data.



Analysis and Results

From the data we have acquired, we propose four questions and make analysis.

Q1: What is the probability of the highest rating horse running first in past seasons (2021-2024)?

Q2: What range of "Win Odds” is suitable for betting in past seasons (2021-2024)?

Q3: What is the average rating (from current rating) of each sire and dam’s sire of the horses currently in Hong Kong Jockey Club?

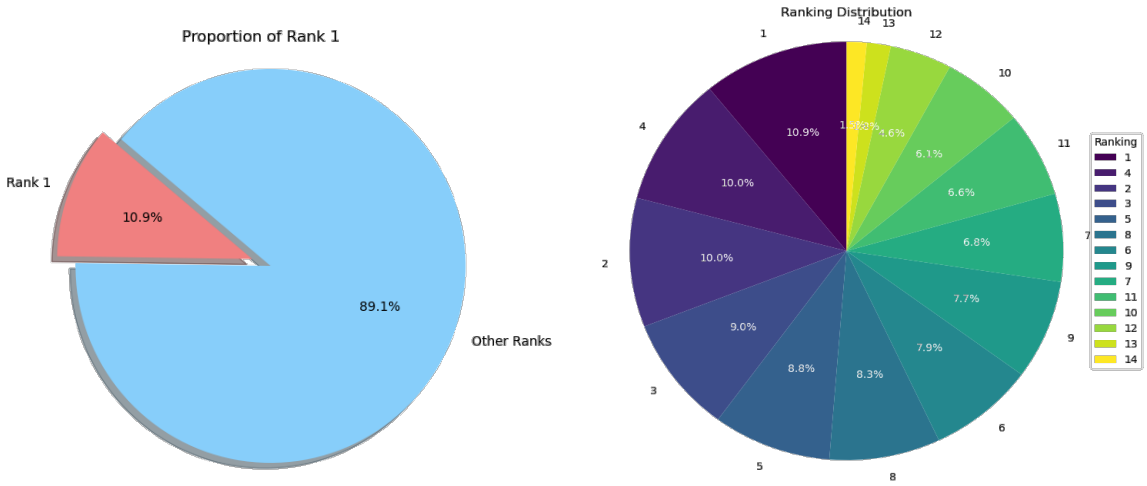
Q4: Which horses are now in the top ten in the Hong Kong Jockey Club’s winning rate （Champion & Top3）?

Result 1

We want to know the performance of the highest rating horses in each race. Is it likely to

always win?

The results are as follows: The probability of champion is not very ideal, but it is normal. Although it seems that the highest rating horse has a low chance of winning the first prize, we can still see that the horses with high ratings are generally ranked relatively high. The ratios for first to fourth place are all around 10%, and the payouts are generally given to the top four, which means that horses with higher ratings can generally meet the public's expectations of them.



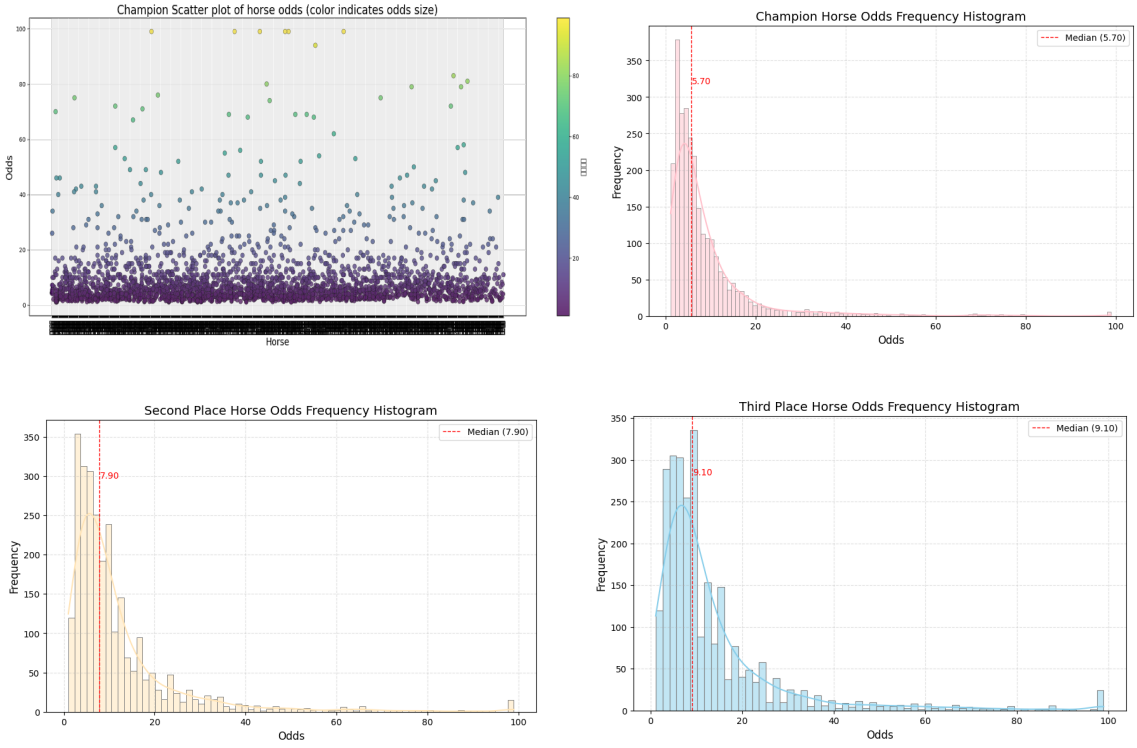
Result 2

We first made a scatter plot of the championship odds and found that the odds were too concentrated, so we made a frequency histogram later.

Since the data distribution is skewed, we marked the median. The champion odds are concentrated at 5.70 or even lower, the second odds are concentrated at 7.90, and the third

place odds are concentrated at 9.10.

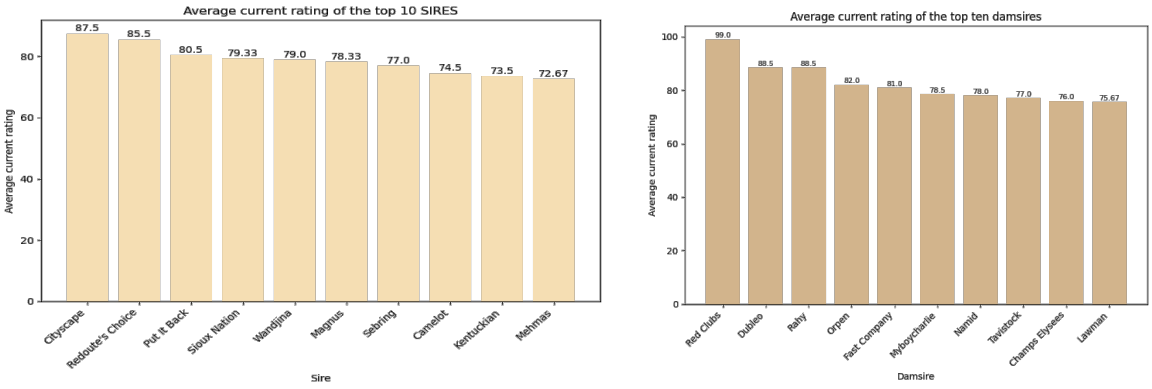
From the visualization we can see that the winning horses generally have lower odds. For example, when we reviewed the competition on December 8, we found that the "浪漫勇士", which had won six games in a row, had its win and position odds even dropped to 1.0-1.1, which was almost a non-profitable state.



Result 3

From the official information, we know that the sire's bloodline has a great influence on the horse's ability, so we try to compare and rank their sire's current ratings and dam's sire's current ratings. We have selected 20 pedigree horses with the highest average scores for the

time being. We can pay attention to the results of horses with the same origin and consider combining this information when betting in the future.



Result 4

The winning rate of a small number of times is unstable, only horses that have competed at least 10 times were selected for analysis. We initially tried to choose more than just three races, but the conditions were still not representative enough, so after comparing 3, 10, and 15 races' data, we finally chose "10" as the dividing point.

```python
import pandas as pd

# 读取处理后的Excel文件
file_path = '/content/cleaned_data.xlsx'
df = pd.read_excel(file_path)

# 清理掉"total_races"大于9的数据
df_cleaned = df[df['total_races'] > 9]

# 保存清理后的数据到新的Excel文件
new_file_path = '/content/cleaned_races.xlsx'
df_cleaned.to_excel(new_file_path, index=False)

print(f"Cleaned data saved to {new_file_path}")
```

```python
# 计算两种胜率，即 冠/总次数 的比率，以及 （冠+亚+季）/总次数 的比率

# 读取处理后的csv文件
file_path = '/content/cleaned_races.csv'
df = pd.read_csv(file_path)

# 计算两种胜率
df['win_rate'] = df['wins'] / df['total_races']
df['top3_rate'] = (df['wins'] + df['seconds'] + df['thirds']) / df['total_races']

# 打印结果或保存到新的csv文件
print(df[['wins', 'seconds', 'thirds', 'total_races', 'win_rate', 'top3_rate']])

# 保存结果到新的csv文件（可选）
new_file_path = '/content/win_rates.csv'
df.to_csv(new_file_path, index=False)
print(f"Win rates saved to {new_file_path}")
```

Because there are two betting ways, "win" and "place", we conducted a winning rate analysis on the current participating horses. That is, we listed the probability of winning the

championship and the probability of getting the top three. The results are shown in the chart. Some horses really stand out.

Top 10 Horses by Win Rate

	马匹名称	win_rate
170	浪漫勇士	0.750000
171	金鑽貴人	0.666667
184	驕陽明駒	0.600000
174	錶之銀河	0.538462
193	合夥奔馳	0.538462
176	永遠美麗	0.500000
172	加州星球	0.464286
39	堅又威	0.416667
47	占士德	0.400000
175	維港智能	0.388889

Top 10 Horses by Top 3 Rate

	马匹名称	top3_rate
215	手機錶霸	1.000000
277	營造組裝	0.909091
170	浪漫勇士	0.900000
184	驕陽明駒	0.900000
171	金鑽貴人	0.875000
172	加州星球	0.857143
190	美麗第一	0.846154
176	永遠美麗	0.833333
2	球星	0.785714
174	錶之銀河	0.769231

Conclusions

This study has meticulously acquired and processed Hong Kong horse racing data from reliable sources, namely the Hong Kong Jockey Club's official website and MaJing. Through a combination of data acquisition methods, including Python libraries like Requests, BeautifulSoup, Pandas, and Selenium, we successfully extracted and cleaned a significant dataset comprising the details and records of many horses. A total of nearly 30,000 pieces of data were obtained. This comprehensive effort allowed us to conduct exploratory analyses, explore horse performance, and provide insights into horse racing trends. For instance, performance of Highest Rating Horses underscores the importance of ratings as a factor in predicting race outcomes but also highlights the variability inherent in horse racing.

However, key predictive factors such as jockey experience, track conditions, and training regimes were not included, which may affect the actual reference significance. Moreover, this research relies on a lot of historical data and may not fully explain the dynamics and real-time changes in horse racing. The static nature of the past results limits their applicability to real-time strategy.

In general, our data may be suitable for novices just as a reference, it is still very limited, and there is no direct guarantee that a bet on a certain horse will win, and we need to build models to further prediction. Maybe machine learning techniques can be applied to enhance prediction accuracy.

Codebook

Hong Kong Jockey Club

Source: https://racing.hkjc.com/racing/information/chinese/Horse/SelectHorse.aspx

Data format：csv

Data collection time：9/11/2024

Table 1: The basic information on the 1,235 horses in Hong Kong.

Variable Name	Description	Data Type	Example
馬匹名称	Name of the horse.	String	安遇
出生地/馬齡	Birthplace of the horse and its age in years.	String	愛爾蘭/5
毛色/性別	The coat color of the horse and its sex.	String	棕/閹
進口類別	Category under which the horse was imported.	String	自購新馬
今季獎金	Prize money earned in the current season.	Numeric	$3,975,000(HKD)
總獎金	Total prize money earned throughout the horse's racing career.	Numeric	$13,993,775(HKD)
冠-亞-季-總出賽次數	Number of wins, places (second), shows (third), and total number of races started.	Numeric	7-2-2-27
最近十個賽馬日出賽場數	The number of races started in the last ten race days.	Numeric	2
現在位置 (到達日期)	Current location of the horse and the date of arrival.	String	Hong Kong (24/09/2024)
進口日期	Date the horse was imported to HK.	Date	16/10/2021
練馬師	Name of the horse's trainer.	String	賀賢

馬主	Name of the horse's owner.	String	王賢訊
現時評分	Current rating of the horse.	Numeric	104
季初評分	Rating of the horse at the beginning of the season.	Numeric	107
父系	Name of the horse's sire (father).	String	Churchill
母系	Name of the horse's dam (mother).	String	Enrol
外祖父	Name of the horse's grandsire (mother's father).	String	Pivotal
同父系馬	List of other horses sired by the same sire.	String	祥勝將軍戰騎飛

Table 2: Recent 3 seasons race records for the above 1,235 horses.

Variable Name	Description	Data Type	Example
马匹名称	The name of the horse participating in the race	String	安遇
场次	The race meeting number	Numeric	243
名次	The finishing position of the horse in the race	Numeric	44
日期	The date of the race	Date	2024/3/24
赛道	The type of track where the race was held	String	沙田草地"A"
途程	The distance of the race (in meters)	Numeric	1200
場地狀況	The condition of the track on the race day	String	好/快
賽事班次	The type or class of the race	String	G1
档位	The starting position of the horse in the race	Numeric	2
評分	The rating score of the horse after the race	Numeric	85
练马师	The name of the trainer responsible for the horse	String	蔡約翰
騎師	The name of the jockey riding the horse	String	麥道朗
頭馬距離	The distance to the winning horse	String, Numeric	“頸位” or “3-3/4”
獨贏賠率	The odds for the horse to win the race	Numeric	2.9
實際負磅	The actual weight carried by the horse during the race (in kilograms)	Numeric	126
沿途走位	The riding position of the horse during the race	Numeric	4 6 6 7 7 4

完成時間	The time taken for the horse to finish the race	String	2.28.11
排位體重	The horse's weight before the race (in kilograms)	Numeric	1072
配备	Equipment used for the horse during the race	String	B2/TT-

Note:

It wasn’t used for data analysis because the retired horse's records were not included, and MaJing's race data was more comprehensive.

MaJing

Data sources: Majing Website (https://racing.on.cc/index.html)

Sample size: 32377(Total); 2705(Championship); 2708(Runner-up); 2700(Quarter); 24264(Loss)

Data format：csv

Data collection time：20/11/2024

Describe: This table records the horse racing data in the last four seasons in Hong Kong. (Incomplete data for 24/25 season) In order to better analyze and organize the data, this table is categorized into 4 sub-tables for Championship(冠), Runner-up(亚), Quarter(季), and Loss (负).

Table 3: Four seasons' horse racing records in Hong Kong.

Variable Name	Description	Data Type	Example
馬匹	The name of the horse participating in the race	String	浪漫勇士
日期	The date of the race	Date	11/17/24
場	The race meeting number	Numeric	190
班次	The type or class of the race	String	G2
路程	The type of track where the race was held and The distance of the race (in meters)	String	田日 B 2000
場地	The condition of the track on the race day	String	好地
分	The rating score of the horse after the race	Numeric	133
配備	Equipment used for the horse during the race	String	T
磅	The actual weight carried by the horse during the race (in kilograms)	Numeric	128
排位重	The horse's weight before the race (in kilograms)	Numeric	1174
騎	The name of the jockey riding the horse	String	麥
練	The name of the trainer responsible for the horse	String	沈
檔	The starting position of the horse in the race	Numeric	1
走位	The riding position of the horse during the race	Numeric	6 5 2
名	The finishing position of the horse in the race	Numeric	1
距離	The distance to the winning horse	String	4 ¼ or 頸
時間	The time taken for the horse to finish the race	Numeric	1.59.70
冠	The name of the champion horse	String	浪漫勇士 or ▲
亚	The name of the runner-up horse	String	嘉應傳承 or ▲
季	The name of the third-place horse	String	觔斗雲 or ▲
頭馬	The time taken for the champion horse to finish the race	Numeric	1.59.70
賠	The win odds of the horse for the game	Numeric	1.1

Note:

(1) Not all horses are equipped, so the “配備” column may be empty.

(2) “▲” in “冠”, “亚” or “季” columns means that the horse in this line get first, second or third place.

(3) Due to data source, trainer and jockey names are incomplete (only family name).

References

Gupta, M. ., & Singh, L. . (2023). Predicting Outcomes of Horse Racing using Machine Learning. International Journal on Recent and Innovation Trends in Computing and Communication, 11(9), 38–47. https://doi.org/10.17762/ijritcc.v11i9.8119

Gupta, M. ., & Singh , L. . (2023). Horse Race Results Prediction Using Machine Learning Algorithms With Feature Selection. International Journal of Intelligent Systems and Applications in Engineering, 12(2s), 132–139. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/3565