

בס"ד  
גֶּל שְׁמוֹאֵל 323844837  
רָחֵל יְהוֹאַלְשֵׁט 207574203

# פרויקט במדעי הנתונים - מרוצי סוסים (דו"ח מסכם)

גֶּל שְׁמוֹאֵל

רָחֵל יְהוֹאַלְשֵׁט



# מבוא

מרוצי סוסים הם אחת מהה פעילויות הספורטיביות העתיקות ביותר בעולם, והם שומרים על מעמדם הפופולרי גם עדין המודרני. סיפור זה, שמתחיל עוד מתקופת רומי העתיקה, עבר שינויים והתפתחות רבות לאורך ההיסטוריה, אך נשאר נאמן למורשתו, כשהוא משלב בין כישרונו הרכיביה, אימון סוסים ברמה גבוהה וידע עמוק על התנהוגות הפיזיולוגית והפסיכולוגית של בעלי חיים אלו.

מעבר להיותם ספורטיבי, מרוצי הסוסים מהווים מרכיב חשוב בתרבותם ובתרבותם רבות, בעיקר במדינות כמו בריטניה, ארצות הברית, אוסטרליה ויפן. בכל אחת מדינות אלו מתקיימים אירועי מרוצי סוסים יוקרתיים, המושכים אליהם קהל רב ומזכירים אתגרים ייחודיים הינם לסוסים והם לרכיבים.

בנוסף, תעשיית מרוצי הסוסים מהוות חלק בלתי נפרד מהכלכלה העולמית, כאשר היא כוללת ענפים רבים כמו הימורים, מסחר בחיות, גידול סוסים גזעים, תעשיית הווטרינריה והאימון, ותרבות עשירה המתפתחת סביב הפעילויות הללו. השוק של מרוצי הסוסים הוא שוק דינמי ובעל נפח כלכלי עצום, כאשר הימורים בלבד מהווים תעשייה בהיקף של מיליארדי דולרים בשנה.

היכולת לחזות את תוצאות המרוצים הפכה לנושא מרכזי בעולם של ניתוח נתונים, כאשר ההשלכות הכלכליות והתרבותיות הן עצומות. בתעשייה זו, כל החלטה קטנה יכולה להשפיע על סכומים גדולים של כסף ועל המוניטין של משתתפים רבים. בזכות השיפורים הטכנולוגיים בעשורים האחרונים, ניתוחים סטטיסטיים ומודלים חכמים מציעים כלים חזקים לחיזוי תוצאות המרוצים. מנתחי נתונים משתניםים במגוון שיטות מתקדמות לשיפור הדיקוק בתוצאות ולהגדלת הרווחיות.

במסגרת פרויקט זה, אנו מתמקדemos בניתוח נתונים היסטוריים של מרוצי סוסים בשנים 1990-2020. המטרה היא לבחון את ההשפעה של שיטות ניתוח מתקדמות על תחומי מרוצי הסוסים ולבדק כיצד ניתן להשתמש במידע זה כדי ליעל את תוצאות התוצאות ואת קבלת החלטות בתחום זה. ניתוח זה יכול בחינת תכונות ומסלולים, ויתמקד במציאת קשרים ומגמות שיכולים לסייע שונות כמו נתונים סוסים בניביות תוצאות המרוצים בצורה מדויקת יותר.

# הבעיה

במטרי סוסים, כמו בספורט מנצח אחר, ההצלחה תלולה במספר רב של משתנים, כגון בריאות הסוס, יכולות הרוכב, תנאי המסלול.

עם זאת, למטרת השפע של הנתונים הזמינים, רבים מההימורים ומקבלי החלטות בתעשייה זו משתמשים על אינטואיציה וניסיון אישי, מה שוביל לרמת חיזוי נמוכה ול失误ון כלכלי גבוה.

האתגר המרכזי בפרויקט זה הוא למצוא דרך לשפר את היכולת לחזות תוצאות מטרדי סוסים בצורה מדויקת יותר על בסיס נתוניים. הפער הקיים כיום הוא חסר במודלים חיזוי מתקדמים שמסוגלים לשקלל את כל המשתנים הרלוונטיים ולהציג תחזיות אמינות. הפער הזה לא רק פוגע ברוחניות של המשתתפים בתעשייה ההימורית, אלא גם מונע שימוש אופטימלי במידע הזמן לצורך שיפור ביצועים והקנת סיכונים.

בפרויקט זה, אנו ננסה לגשר על פער זה באמצעות שימוש בשיטות מתקדמות של ניתוח נתונים, כמו מודל עץ החלטה וניתוחים סטטיסטיים. המטרה היא לפתח מודל חיזוי שמסוגל לקחת בחשבון את המורכבות של המטרדים ולספק תחזיות מדויקות יותר שיכולים להוביל לקבלת החלטות מושכלת יותר, הן עבור המתחרים והן עבור ההימורים.

# השיטה

בהתאם לאתגרים בניתוח תוצאות מרוצי סוסים, נעשה שימוש בשיטות ניתוח מתקדמות ובמגון כלים וספריות בפייתון. העבודה הchallenge בהכנות הנתונים, המשיכת בניתוח ראשוןי, והתקדמה לניתוח עמוק.

הכנות הנתונים:

בשלב הראשון של העבודה, ביצענו אחזור והכנה של נתונים המרוצים ממגרי מידע חיצוניים. השתמשנו ב-API Kaggle להורדת מערכិ נתוני המכילים מידע על סוסים ומרוצים משנת 1990 עד שנת 2020, נזרנו בספריות os ו-chosj לטעינה ובספריית

usezip כדי לחלץ את תוכנו לתוך התקינה המעודת. לאחר מכן איחוד נתונים הסוסים עבור כל שנה ל-DataFrame בשם

## horses\_data

	rid	horseName	age	isFav	trainerName	jockeyName	position	decimalPrice	saddle	positionL	weight
0	271018	Combermere	6.0	0	R G Frost	J Frost	1	0.222222	0.0	NaN	69
1	271018	Royal Battery	6.0	0	D H Barons	S Earle	2	0.090909	0.0	10	69
2	271018	Just So	7.0	0	J D Roberts	S Burrough	3	0.029412	0.0	15	66
3	271018	Mandraki Shuffle	8.0	0	Oliver Sherwood	M Richards	4	0.090909	0.0	20	69
4	271018	Turnberry Dawn	8.0	0	T B Hallett	P Richards	5	0.047619	0.0	dist	69
...	...	...	...	...	...	...	...	...	...	...	...
4107310	415090	Beefeater	8.0	0	Roydon Bergerson	Hazel Schofer	7	0.030303	6.0	shd	58
4107311	415090	Aimee's Jewel	4.0	0	Trudy Keegan	Lisa Allpress	8	0.153846	11.0	.5	57
4107312	415090	Times Ticking	5.0	0	Alby Macgregor	Jonathan Riddell	9	0.044053	8.0	hd	58
4107313	415090	Shadows Cast	8.0	0	Mark Oulaghan	Johnathan Parkes	10	0.041152	3.0	1	58
4107314	415090	Awesome Al	7.0	0	Buddy Lammas	Ryan Bishop	11	0.016393	12.0	2.75	57

וקוד לאיחוד נתונים המרוצים (races) DataFrame בשם .races\_data

	rid	course	time	title	ages	date	distance	countryCode	condition	hurdles	winningTime	distanceKM
0	271018	Exeter	03:15	David Garrett Memorial Challenge Trophy Novice...	6-8yo	1990	3m1f	GB	Soft	19 fences	398.30	5.029187
1	275156	Tramore (IRE)	02:00	Tattersalls Mares E.B.F. Novice Chase	NaN	1990	2m	IE	Soft	12 fences	267.00	3.218680
2	282203	Catterick	02:45	Scotch Corner Handicap Chase	NaN	1990	1m7½f	GB	Good To Firm	12 fences	238.00	3.118096
3	298761	Cheltenham	02:30	A. S. W. Handicap Hurdle	NaN	1990	2m	GB	Good To Firm	NaN	243.80	3.218680
4	301118	Windsor	03:30	Touchen End Handicap Hurdle	NaN	1990	2m6f	GB	Good	NaN	330.70	4.425685
...	...	...	...	...	...	...	...	...	...	...	...	...
396567	415086	Hanshin (JPN)	06:45	Challenge Cup (Grade 3) (3yo+) (Turf)	3yo+	2020	1m2f	JP	Firm	NaN	119.90	2.011675
396568	415087	Los Alamitos (USA)	10:30	Starlet Stakes (Grade 1) (2yo Fillies) (Main T...	2yo	2020	1m½f	US	Fast	NaN	104.53	1.609340
396569	415088	Nakayama (JPN)	06:25	Sports Nippon Sho Stayers Stakes (Grade 2) (3y...	3yo+	2020	2m2f	JP	Good	NaN	232.00	3.621015
396570	415089	Tampa Bay Downs (USA)	05:10	Maiden Claiming Race (2yo) (Turf)	2yo	2020	1m	US	Firm	NaN	99.34	1.609340
396571	415090	Trentham (NZ)	03:45	Rydges Wellington Captain Cook Stakes (Group 1...	2yo+	2020	1m	NZ	Soft	NaN	99.66	1.609340

ולסיום מיזגנו את `a` DataFrame-הן `races_data` ו- `horses_data` בשם `merged_data`.

rid	horseName	age	isFav	trainerName	jockeyName	position	decimalPrice	saddle	positionL	...	time	title	ages	date	distance	countryCode	condition	hurdles	winningTime	distanceRM	
0	271018	Combermere	6.0	0	R G Frost	J Frost	1	0.222222	0.0	NaN	...	03:15	David Garrett Memorial Challenge Trophy Novice...	6-8yo	1990	3m1f	GB	Soft	19 fences	398.30	5.029187
1	271018	Royal Battery	6.0	0	D H Barons	S Earle	2	0.090909	0.0	10	...	03:15	David Garrett Memorial Challenge Trophy Novice...	6-8yo	1990	3m1f	GB	Soft	19 fences	398.30	5.029187
2	271018	Just So	7.0	0	J D Roberts	S Burrough	3	0.029412	0.0	15	...	03:15	David Garrett Memorial Challenge Trophy Novice...	6-8yo	1990	3m1f	GB	Soft	19 fences	398.30	5.029187
3	271018	Mandrake Shuffle	8.0	0	Oliver Sherwood	M Richards	4	0.090909	0.0	20	...	03:15	David Garrett Memorial Challenge Trophy Novice...	6-8yo	1990	3m1f	GB	Soft	19 fences	398.30	5.029187
4	271018	Tumberry Dawn	8.0	0	T B Hallett	P Richards	5	0.047619	0.0	dist	...	03:15	David Garrett Memorial Challenge Trophy Novice...	6-8yo	1990	3m1f	GB	Soft	19 fences	398.30	5.029187
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
#107310	415090	Beefeater	8.0	0	Roydon Bergerson	Hazel Schofer	7	0.030303	6.0	shd	...	03:45	Rydges Wellington Captain Cook Stakes (Group 1...)	2yo+	2020	1m	NZ	Soft	NaN	99.66	1.609340
#107311	415090	Aimee's Jewel	4.0	0	Trudy Keegan	Lisa Allpress	8	0.153846	11.0	.5	...	03:45	Rydges Wellington Captain Cook Stakes (Group 1...)	2yo+	2020	1m	NZ	Soft	NaN	99.66	1.609340
#107312	415090	Times Ticking	5.0	0	Alby Macgregor	Jonathan Riddell	9	0.044053	8.0	hd	...	03:45	Rydges Wellington Captain Cook Stakes (Group 1...)	2yo+	2020	1m	NZ	Soft	NaN	99.66	1.609340
#107313	415090	Shadows Cast	8.0	0	Mark Oulaghan	Johnathan Parkes	10	0.041152	3.0	1	...	03:45	Rydges Wellington Captain Cook Stakes (Group 1...)	2yo+	2020	1m	NZ	Soft	NaN	99.66	1.609340
#107314	415090	Awesome Al	7.0	0	Buddy Lammes	Ryan Bishop	11	0.016393	12.0	2.75	...	03:45	Rydges Wellington Captain Cook Stakes (Group 1...)	2yo+	2020	1m	NZ	Soft	NaN	99.66	1.609340

### שיטת העבודה וניתוח ראשוני:

לאחר הכתנת הנתונים, התחלנו בניתוח ראשוני של הדאטאסטים. לצורך עיבוד וניתוח הנתונים השתמשנו בספריית **pandas** ליצירת טבלאות לחקר מגמות ראשוניות, תוך שימוש במגוון שיטות לעיבוד נתונים כמו מינון, קיבוץ ויצירת טבלאות ציר.

לאחר עיבוד הנתונים הראשוני, עברנו לשלב ההדמיה וניתוח עמוק של הנתונים, תוך שימוש במספר ספריות מתקדמות:

- **Numpy**: שימשה אותנו לביצוע חישובים מתמטיים כגון חישוב ממוצעים, סטיות תקן ומינימופלים על מערכים.
- **Seaborn**: אפשרה ייצור גרפים ויזואליים כמו **heatmaps** לזהוי דפוסים ומטריצות קוורלציה.
- **Plotly ו- Matplotlib**: להציג קשרים בין משתנים מרכזיים ולחקר הנתונים.
- **Matplotlib**: אוניברסיטאיyi כמו ייצור היסטוגרמות, דיאגרמות פיזור ו-**line charts**.
- **Scikit-learn**: שימשה לפיצול הנתונים ולהכרשת מודלים כמו רגסיה לינארית ועכ החלטה, והערכנו את ביצועי המודלים באמצעות מדדים שונים.



### **ניתוח עמוק:**

בשלב הניתוח המעמיק, עברנו להערכתה של קשרים מורכבים יותר בין משתנים שונים באמצעות טכניקות מתقدמות. בינו מודלים של רגסיה לינארית לצורך חיזוי תוצאות המרוצים, תוך התחשבות במשתנים כמו גיל הסוס, משקלו ותנאי המסלול. התוצאות הצביעו על קשר מובהק בין גיל הסוס לבין הזמן שבו הוא סיים את המירוץ.

בנוסף, השתמשנו בעז החלטה לצורך ניתוח הסתברותי של ניצחונות הסוסים. המודל מאפשר לנו לzechות את הגורמים המשפיעים ביותר על הסיכוי לניצחון, כגון גיל הסוס ותנאי המסלול. ראיינו שקיים קשר חזק בין תנאי המסלול לבין סיכוי לניצחון, כאשר תנאים טובים יותר העלו משמעותית את הסיכויים לניצחון.

כדי להמחיש את התוצאות ולהזכיר את הנתונים באופן אינטראקטיבי, השתמשנו בספרייה **Plotly** ליצור גרפים תלת-ממדיים שהתמקדו בקשרים בין גיל הסוס, משקלו והמקום שבו הוא הגיע. הדרימות האינטראקטיביות הללו אפשרו לנו להביע בচורה ויזואלית את התפלגות הנתונים ואת הקשרים המרכזיים בין המשתנים.

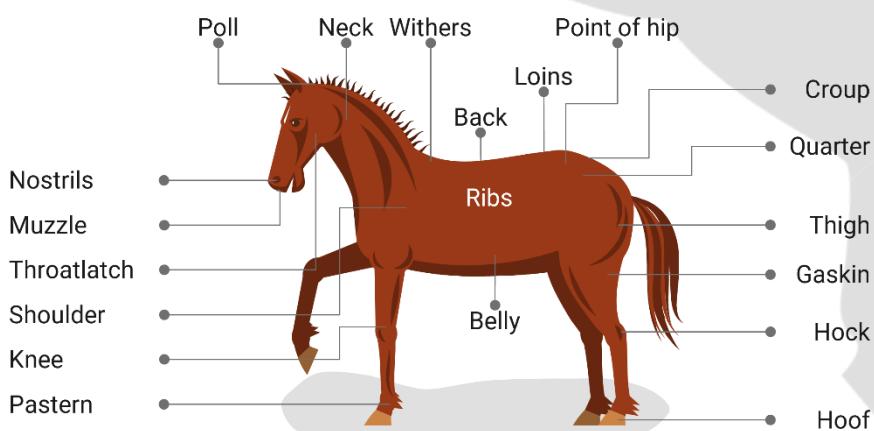
# תיאור הדאטאסטים

בפרויקט זה נעשה שימוש בדאטאסט שנקרא horse-racing שנלקח ממאגרי מידע מאתר Kaggle, המכיל מידע מפורט על מרוצי סוסים שנערכו בין השנים 1990 ל-2020.

הדאטאסט מחולק לקבצים של horses ולקבצים של races עבור כל שנה. הקבצים של horses כוללים מידע על הסוסים המשתתפים במרוצים לאורך השנים, לפרוייקט שלנו התמקדנו בעמודות שנראו לנו הכי רלוונטיות כמו:

- **Rid**: מספר מזהה של המירוץ.
- **HorseName**: שם הסוס.
- **Age**: גיל הסוס במועד המירוץ.
- **IsFav**: משתנה בינאריה המציין אם הסוס היה פיבורייט במירוץ (1: פיבורייט, 0: לא פיבורייט).
- **TrainerName**: שם המאמן.
- **JockeyName**: שם הרוכב.
- **Position**: המיקום שבו הסוס סיים את המירוץ (אם הסוס סיים במקום 40 ז"א שהוא לא סיים את המסלול).
- **DecimalPrice**: רווח זכיה על כל סכום מطبع שהכנסת.
- **Saddle**: מכשולים.
- **PositionL**: מרחק הסוס מהסוס שלפניו.
- **Weight**: משקל הסוס.

כדי שלא נצטרך לעבור על כל שנה בנפרד בניתוח כתבנו קוד לאיחוד הנתונים horses שכלל טעינה של קבצי CSV עבור כל שנה בנפרד, מיצוי העמודות הרלוונטיות, והוספת הנתונים ל-DataFrame בשם **horses\_data**.



הकבצים של races מכילים נתונים שנערכו בין השנים 1990-2020, הנתונים על המרוצים עצם, מהדאטאסט המצוין לקחנו את העמודות העיקריות הבאות:

- :Rid – מזהה המירוץ.
- :Course – מסלול.
- :Time – זמן המירוץ.
- :Title – שם המירוץ.
- :Ages – מאיזה גיל אפשר להתחרות במסלול.
- :Distance – אורך המסלול.
- :CountryCode – איפה המסלול התקיים.
- :Condition – תנאי המירוץ.
- :Hurdles – כמות המכשולים שייהיו במסלול.
- :WinningTime – משך הזמן שהקח לסיים במקום הראשון לניצחון.

גם עבור races כתבנו קוד לאיחוד הנתונים שככל טעינה של קבצי ה-CSV לכל שנה בנפרד, מיצוי העמודות הרלוונטיות, הוספה עמודה חדשה שנקראת **distanceKM** שבאה מופיע אורך המסלול ב-**KM** לאחר המرة שנעשתה באמצעות פונקציה שכתבנו שנקראת **convert\_distance**, בנוסף, עקב ידעה שנאחד את 30 הדאטאסטים יצרנו עמודה נוספת נוספת שתוכל לידע אותנו את שנת מסלול המירוץ בשם **Date** לאחר כל התוספות הללו איחדנו את הנתונים ל-**DataFrame** בשם **races\_data**.

בנוסף יצרנו **merged\_data** בשם **DataFrame** שהוא בעצם מיזוג של **races\_data & horses\_data**, שמכיל את המידע גם על הסוסים וגם על המרוצים שבהם השתתפו, מה שמאפשר ניתוח מקיף יותר של הנתונים.

# הסבר לתהילך הממחקר:

תהליך הממחקר שלנו התבסס על גישה רב-שלבית, שכלה את השלבים הבאים:

## 1. איסוף הנתונים:

התחלנו את הפרויקט באיסוף נתונים ממגרי מידע חיצוניים עבור מרצוי סוסים מהשנים 1990-2020, תוך שימוש ב-**Kaggle API Dataset** שנקרא **horse-racing** ובו קבצים המכילים מידע על הסוסים ועל המרצים. בעזרת הספריות **os** ו-**json**, טענו את פרטי ההתחרויות ל-Kaggle מתוךקובץ **JSON** ששמר את שם המשתמש והמפתח (**API key**). השתמשנו בפרטים אלו כדי להציג את משתני הסביבה הנדרשים לביצוע בקשות ל-**Kaggle API**. לאחר מכן, השתמשנו ב-**Kaggle API Dataset** להורדת ה-**Dataset** של מרצוי הסוסים, ועשינו שימוש בספריית **zipfile** כדי לחלק את תוכנו לתוכה התקינה המיעודת. וידאנו את ההצלחה של הצלחן בחילוץ באמצעות הדפסה של רשימת הקבצים שהתקבלו.

## 2. איחוד הדאטה והכנתו לניתוח:

לאחר איסוף הנתונים, ביצענו איחוד עבור הקבצים של **Horses** ועבור הקבצים של **Races** באמצעות קוד לאיחוד כדי ליצור דאטה מסודר שאפשר לנתח בצורה קלה ונוחה. הקוד של האיחוד עבר כל קבוצת קבצים כלל טעינה של קבצי CSV עבור כל שנה בנפרד, מציו העמודות הרלוונטיות, והוספה הנתונים ל-**DataFrame** אחד. עבור **horses** קראנו ל-**DataFrame** בשם **horses\_data** ועבור **races** קראנו ל-**DataFrame** בשם **races\_data**. בשלב זה, ביצענו פעולות ניקוי ועיבוד ראשוני, כמו הרמת ייחדות מדידה (לדוגמה, ממיל לקלומטר), זיהוי ערכים חסרים וטיפול בהם, ושמירה על עמודות חשובות בלבד לצורכי ניתוח.

לאחר מכן, בעזרת קוד איחוד שכלל התאמת מזהי מירוץ (**rid**) ביצענו איחוד של המידע של הסוסים (**horses\_data**) עם המידע של המרצים (**races\_data**) ליצירת מערך נתונים מאוחד בשם **merged\_data** זהו מערך נתונים שמכיל מידע מדויק על הסוסים, הרוכבים, המסלולים, ותנאי המירוץ, ובכך מאפשר לנו לחקור את הקשרים האפשריים בין משתנים רבים.

### 3. ניתוח ראשוני

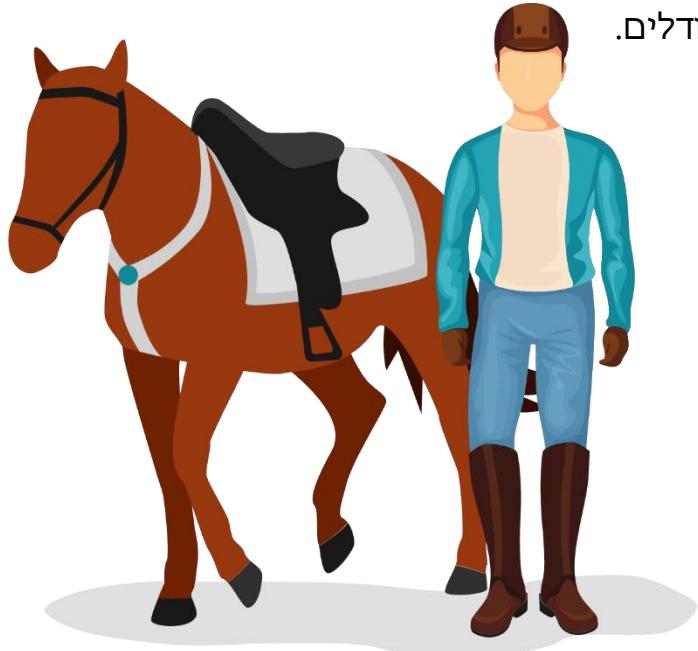
בשלב הניתוח הראשוני, השתמשנו בכלים כמו **Pandas** ו-**Seaborn** לחקירת המוגמות הבסיסיות בDATA. חקרנו את הקשרים הבסיסיים בין משתנים כמו גיל הסוס והמשקל שלו לבין המיקום שבו סים את המירוץ. יצרנו טבלאות ציר וגרפים ויזואליים (כמו **scatter\_3d** ו-**Histogram**) על מנת לקבל תמונה ראשונית של ההתפלגיות והקורלציות בDATA.

### 4. בניית מודלים לחיזוי

בשלב הבא, התמקדנו בבנייה מודלים לחיזוי תוצאות המרוצים. השתמשנו במודלים שונים כמו **Decision Trees** ו-**Linear regression**. כל מודל עבר תהליך של אימון, בדיקה והערכתה, תוך שימוש במשתנים חשובים כמו גיל הסוס, משקלו, ותנאי המסלול. באמצעות המודלים, ניסינו לחזות את מיקום הסוס במירוץ וכן לבחון את ההשפעה של המשתנים השונים על סיכוי לניצחון.

### 5. הדמיה מתקדמת

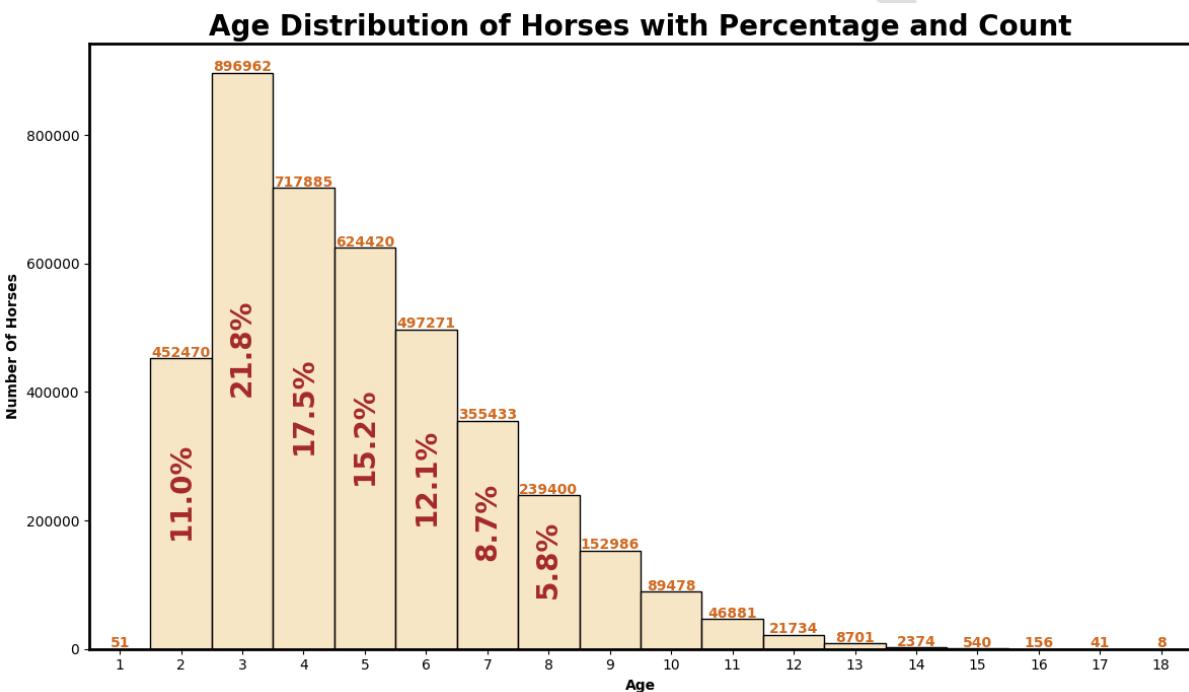
לאחר בניית המודלים, השתמשנו בספריות הדמיה כמו **sklearn** ו-**Plotly** - ליצירת גרפים אינטראקטיביים שהתמקדו בקשרים המורכבים בין משתנים. הדמויות אלו עזרו לנו להבין באופן ויזואלי את הקשרים המרכזיים בין המשתנים ולגלות דפוסים נסתרים בDATA. תהליך המחקר לווה לאורך כל הדרך בבדיקות תיקוף והערכתה, כאשר נבדקו ביצועי המודלים באמצעות מדדים סטטיסטיים שונים על מנת להבטיח את איכות התחזיות והמודלים.



# תוצאות:

## ניתוח התפלגות גיל הסוסים והשפעתו על השתתפות במרוצים:

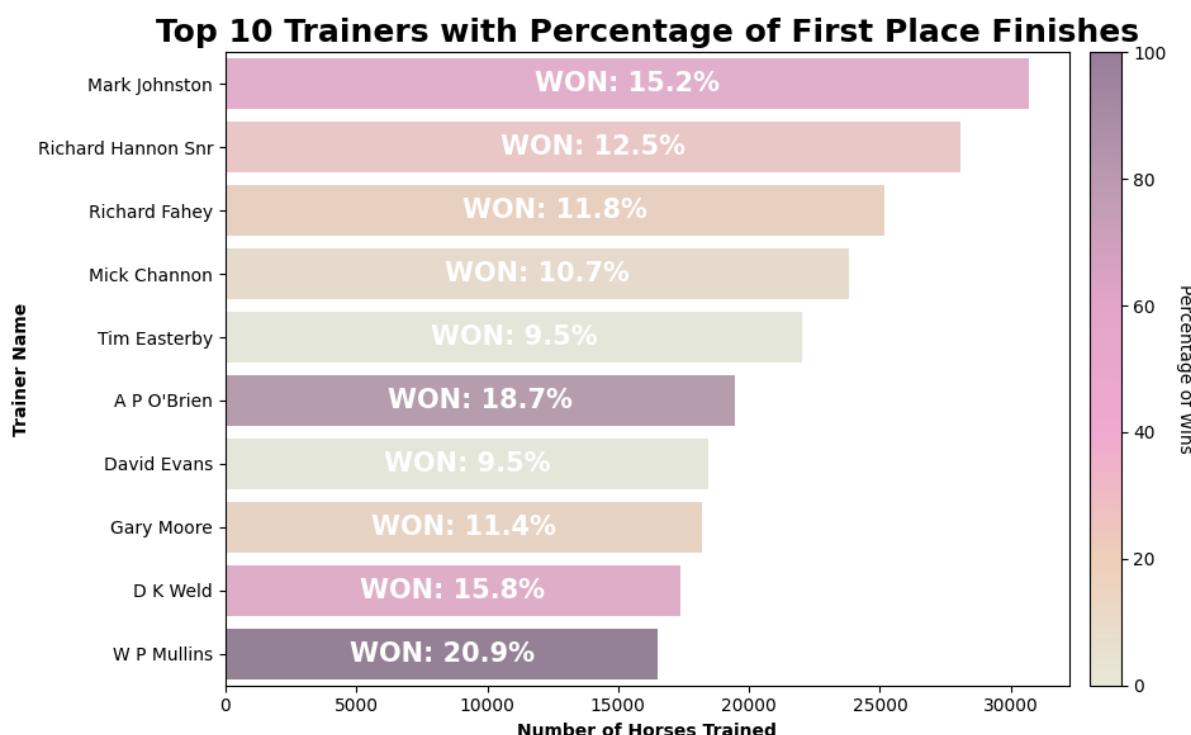
הגרף מתאר את התפלגות הגילאים של סוסים שהשתתפו במרוצים לאורך השנים. מהגרף ניתן לראות כי רוב הסוסים המתחרים הם בגילאי 3 עד 6 שנים. גיל 3 הוא הנפוץ ביותר, עם 896,962 סוסים שהם כ- 21.8% מכלל הסוסים במאגר הנתונים. הגילאים 4 ו-5 גם הם מייצגים אחוזים משמעותיים, עם 17.5% ו- 15.2% בהתאמה. לאחר מכן, חלה ירידה מתמדת במספר הסוסים ככל שהגיל עולה, כאשר בגילאים מעל 10 הכמות הולכת ופוחתת בצורה ניכרת, ומעל גיל 15 כמעט אין סוסים במאגר.



- התפלגות זו מעידה על כך שסוסים בגילאים צעירים יותר נוטים להשתתף יותר במרוצים, ככל הנראה בשל יכולות הפיזית והמנטלית הטובה יותר בהשוואה לסוסים מבוגרים יותר.

### ניתוח אחוז ניצחונות של עשרת המאמנים המובילים:

הגרף מציג את עשרת המאמנים המובילים במספר הסוסים שהם אימנו, יחד עם אחוז הניצחונות שלהם. מהנתונים ניתן לראות כי המאמן עם אחוז הניצחונות הגבוה ביותר הוא **W P Mullins** עם 20.9% ניצחונות, לעומת זאת שמשיר הניצחונות השני הוא **Mark Johnston**, המאמן שאימן את הסוסים הגבוהים ביותר ביותר של סוסים (מעל 30,000), הגיע רק לאחוז ניצחונות של 15.2%.



- ישנה מגמה כללית שבה מאלפים עם מספר גובה של סוסים מאמנים נוטים לקבל אחוז זכייה מעט נמוך יותר, אולי בשל האתגרים של ניהול מרבית סוסים.
- הגרף מדגיש למעשה את האיזון בין כמות (מספר סוסים מאמנים) וaicות (אחוז ניצחון) בקרב המאמנים המובילים.

## ניתוח מגמות ביצוע של חמישת המאמנים המובילים לאורך השנים:

### מגמות כלליות לאורך השנים:

המגמות העיקריות שנראות בגרף מצביעות על כך שככל אחד מחמשת המאמנים המובילים עבר תנודות משמעותיות במספר הניצחונות והמרוצים לאורך השנים. ניתן לראות שמאנו מסיים עשוי לבנותו במיוחד באחת השנים, וזה לאבד את יתרונו בשנים אחרות.

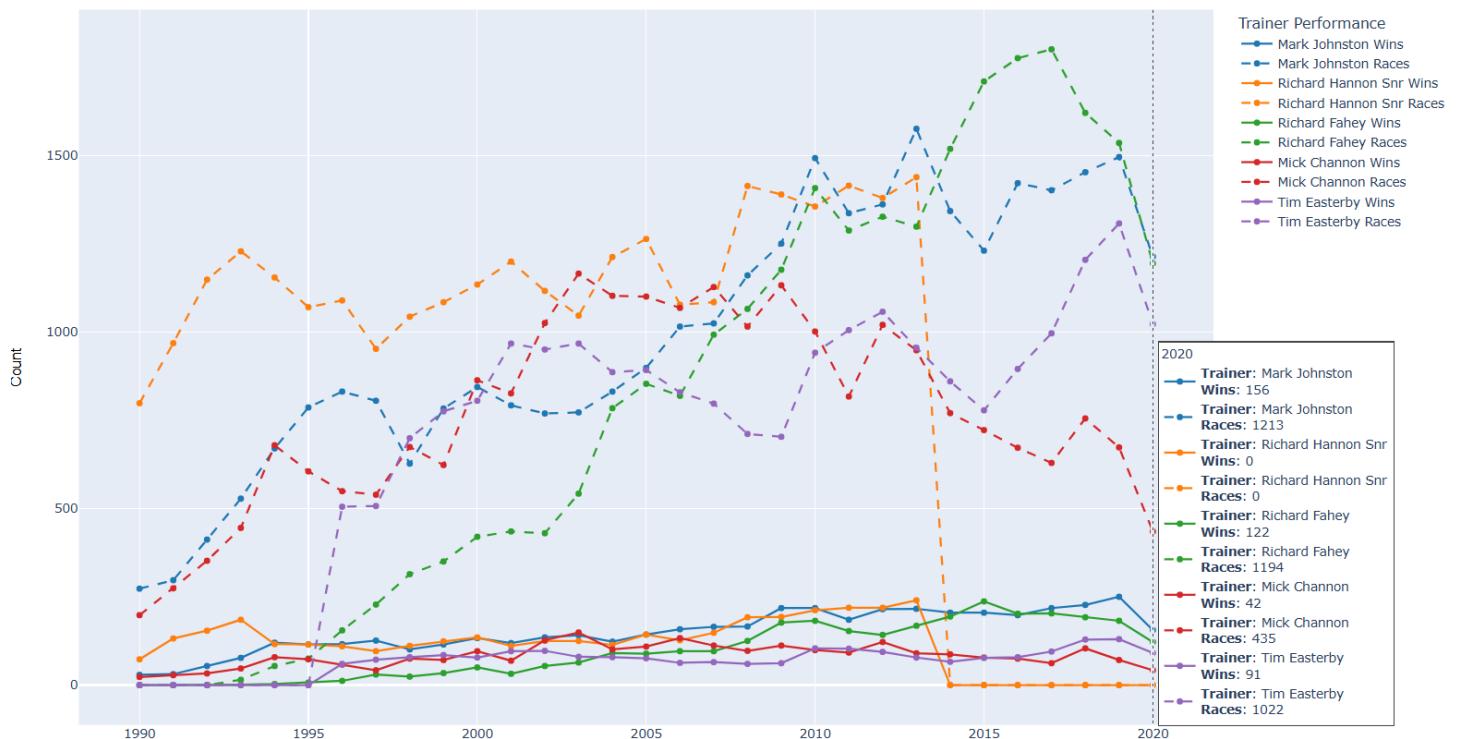
**Richard Hannon Snr:** מציג מגמת עלייה בולטת בניצחונות עד לשיא בסביבות 2013, אך לאחר מכן יש ירידה חדה במספר הניצחונות והמרוצים.

**Mark Johnston:** שומר על יציבות יחסית לאורך השנים, אך גם הוא חווה תנודות קלות בניצחונות ובמספר המרוצים.

**Richard Fahey ו-Mick Channon:** מציגים מגמות משתנות אך שומרות על רציפות מסוימת במספר הניצחונות והמרוצים.

- ניתן לראות התאמה כללית בין מספר המרוצים שמאנו מסיים משתתף בהם לבין מספר הניצחונות שלו, אך לא תמיד מדובר בתאמה מדויקת. לדוגמה, בשנים האחרונות ישנו מרוץים רבים אך מספר הניצחונות נמור יחסית.

Wins and Races Over the Years for Top 5 Trainers



## ניתוח אוחחי הניצחונות של מאמנים מובילים בתנאי מירוץ שונים:

אוחחי ניצחונות בתנאי מירוץ איטיים (**Slow**):

המאמן "Gary Moore" השיג את אוחח הניצחונות הגבוה ביותר ביוור (66.7%) בתנאי מירוץ איטיים.

תנאי מירוץ קשים (**Hard**):

רוב המאמנים אינם זוכים לניצחונות בתנאים קשים, מה שמתבטא בערך "XXX" המופיע ברוב המאמנים עבור תנאי מירוץ אלו.

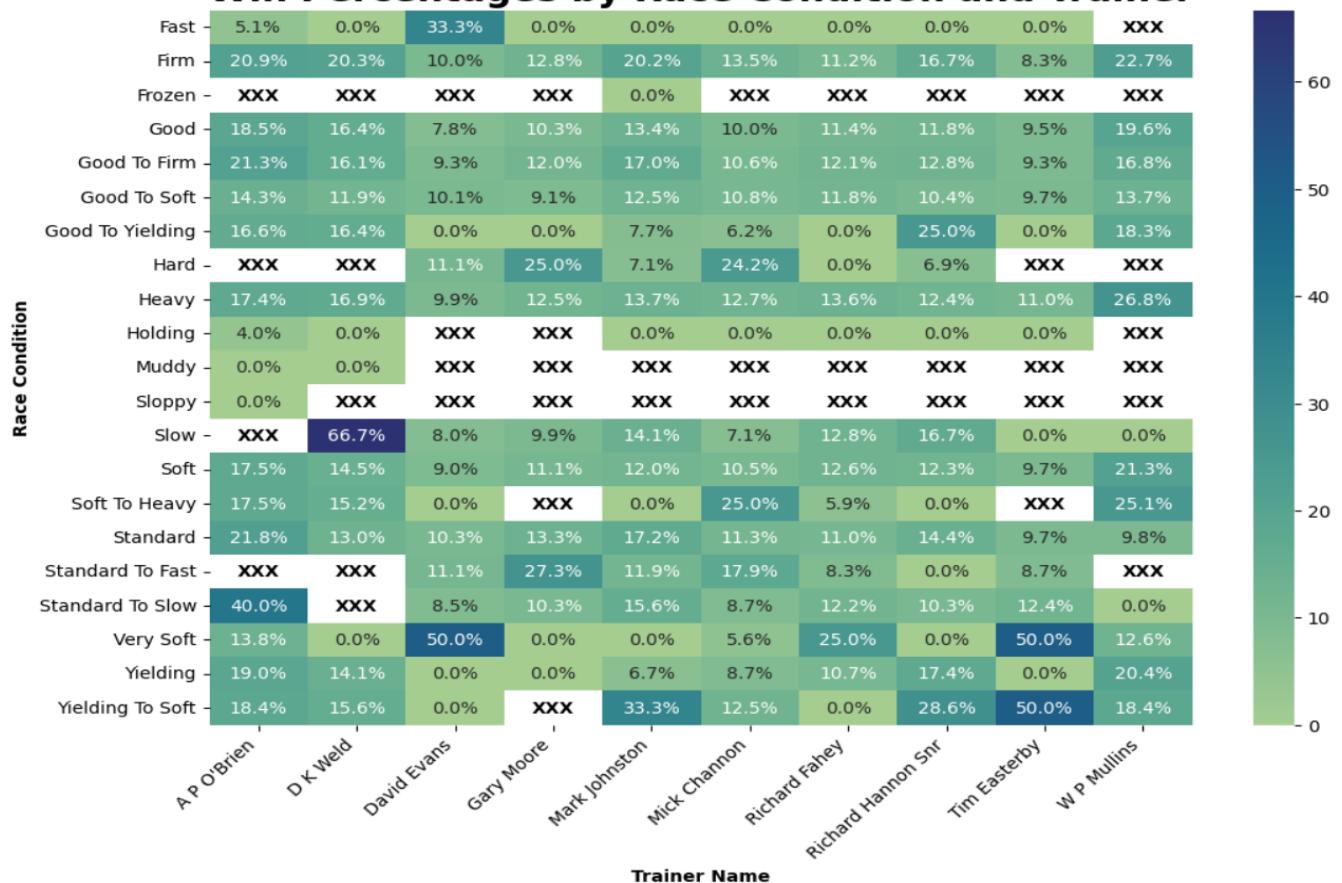
אוחחי ניצחונות בתנאי מירוץ סטנדרטיים (**Standard To Slow**):

המאמן "Tim Easterby" השיג 50% אוחחי ניצחון בתנאי מירוץ סטנדרטיים איטיים.

שונות באוחחי ניצחונות של מאמנים אחרים:

מאמנים אחרים מציגים אוחחי ניצחונות שונים בתנאים שונים, עם תאים רבים המראים את הערך "XXX" בשל חוסר נתונים או השתתפות באותו תנאי מירוץ.

**Win Percentages by Race Condition and Trainer**



הגרף ממחיש שביצועי המאמן תלויים מאוד בתנאי המירוץ. נראה כי כמה מאמנים, כמו David Evans ו-W P Mullins, מתחמים בתנאים מסוימים, ומציגים שיעורי זכייה גבוהים משמעותית. מצד שני, נראה שמצבים מסוימים כדוגמת **Frozen**, **Sloopy** ו-**Muddy** נמנעים או פחות מוצלחים עבור רוב המאמנים. הבנת הקשר בין ביצועי מאמן ותנאי מירוץ יכולה להיות יתרון אסטרטגי לביצוע תוצאות במרוצי סוסים.

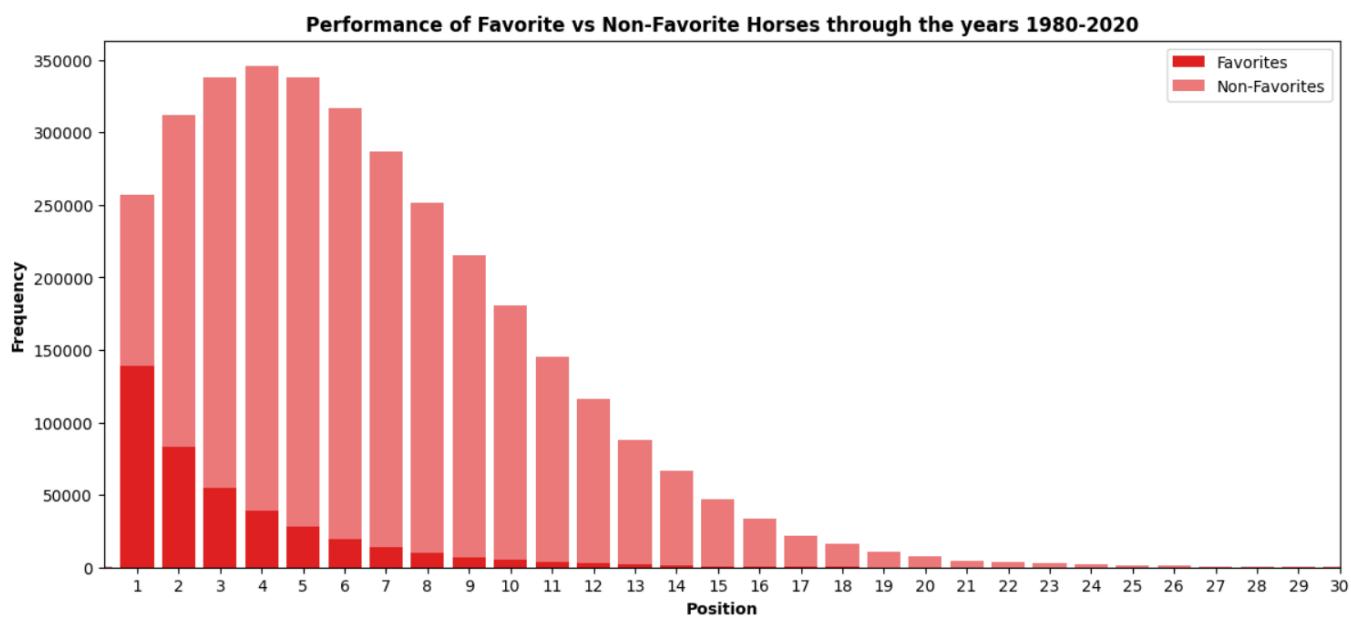
### ניתוח ביצועי סוסים פיבוריטים מול סוסים לא פיבוריטים:

#### סוסים פיבוריטים:

הסוסים הפיבוריטים (**Favorites**) מציגים שכיחות גבואה יותר במקומות הראשונים במרוצים, במיוחד במקומות 1 עד 5. המיקום הראשון, שהוא הניצחון, מציג שכיחות נמוכה יותר בהשוואה למיקומים 2-5, אך עדין מראה שכיחות משמעותית.

#### סוסים לא פיבוריטים:

הסוסים הלא פיבוריטים (**Non-Favorites**) מופיעים בכמותות גדולות יותר במקומות 4 ומעלה, עם שכיחות גבואה במיוחד במקומות 5-10. במקום הראשון, הסוסים הלא פיבוריטים מופיעים בשכיחות נמוכה יחסית.

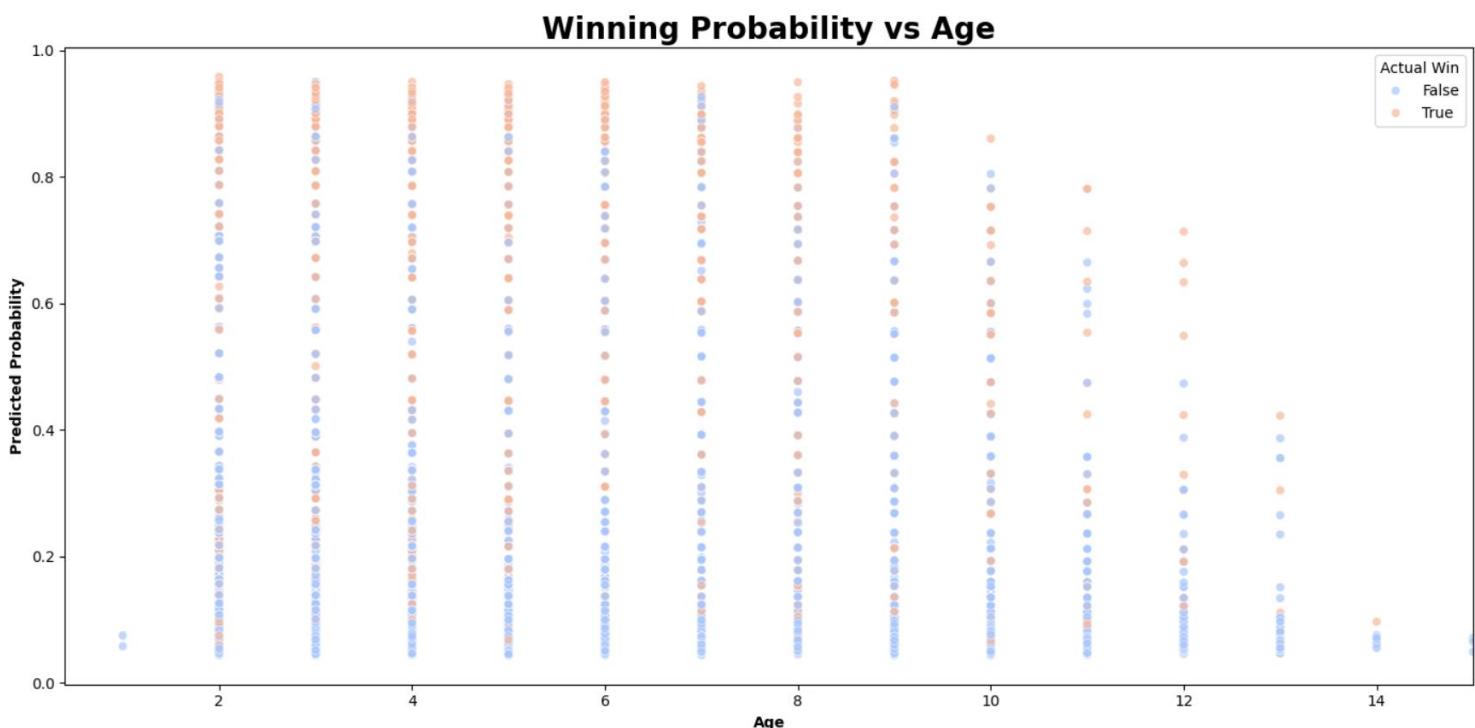


- הגרף מראה שהפיבוריטים מצליחים יותר ו מגעים לעתים קרובות יותר במקומות גבויים, בעוד שסוסים לא פיבוריטים ממוקמים בעיקר במקומות הנמוכים.

## ניתוח הסתברות לניצחון של סוסים לפי גיל: (תוצאות מודל גרסיה לוגיסטי)

הגרף מתאר את ההסתברות החזויה לניצחון של סוסים במירוץ כפונקציה של גילם, תוך שימוש במודל גרסיה לוגיסטי. הגרף מראה את ההסתברות החזויה לניצחון (ציר ה-Y) ביחס לגיל הסוס (ציר ה-X), כשהצבעים השונים מציננים אם הסוס אכן ניצח (True) או לא ניצח (False) במירוץ בפועל.

נראה כי הסוסים בגילאי 5-2 הם בעלי ההסתברות הגבוהה ביותר לניצחון, כפי שנראה מהתקבצותם של נקודות עם הסתברות גבוהה באזוריים אלו. מעבר לגיל 8, ההסתברות לניצחון הולכת ו יורדת, כאשר הסוסים המבוגרים יותר (גילאים 10 ומעלה) מציגים הסתברויות נמוכות יותר לניצחון. ישנה שונות משמעותית בהסתברות לניצחון עבור סוסים בכל גיל, אך המגמה הכללית מראה על ירידת בהסתברות לניצחון עם העלייה בגיל.

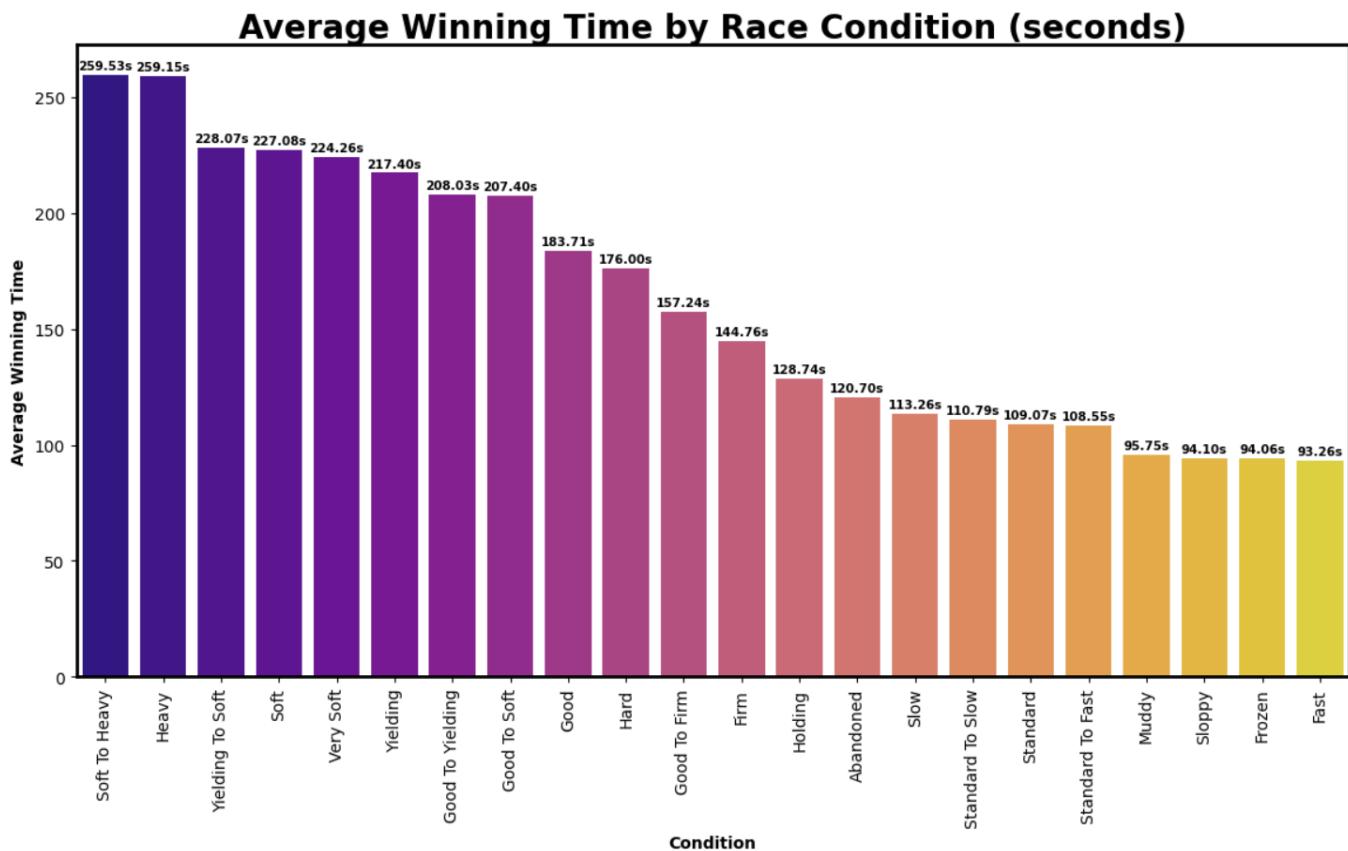


- הגרף מראה שההסתברות לניצחון היא הגבוהה ביותר עבור סוסים בגילאי 2-5 ו יורדת באופן כללי עם העלייה בגיל, במיוחד לאחר גיל 8.

### תוצאות ניתוח זמן זכיה ממוצע לפי מסלול:

הגרף המוצג מציג את זמן הזכיה הממוצע לפי מסלול במרוץ סוסים, בו ערכו הזמן נמדדים בשניות. הזמן הממוצעים חושבו לכל אחד ממצבי המסלול הקיימים בתנונות, ולאחר מכן הגרף ממון בסדר יורד.

ניתן לראות שמצבי מסלול "Heavy" ו-"Soft To Heavy"- מתאפיינים בזמן זכיה ממוצעים ארוכים במיוחד (מעל 250 שניות), בעוד שמצבי מסלול "Fast" ו-"Frozen"- מתאפיינים בזמן זכיה ממוצעים קצרים (פחות מ-100 שניות).

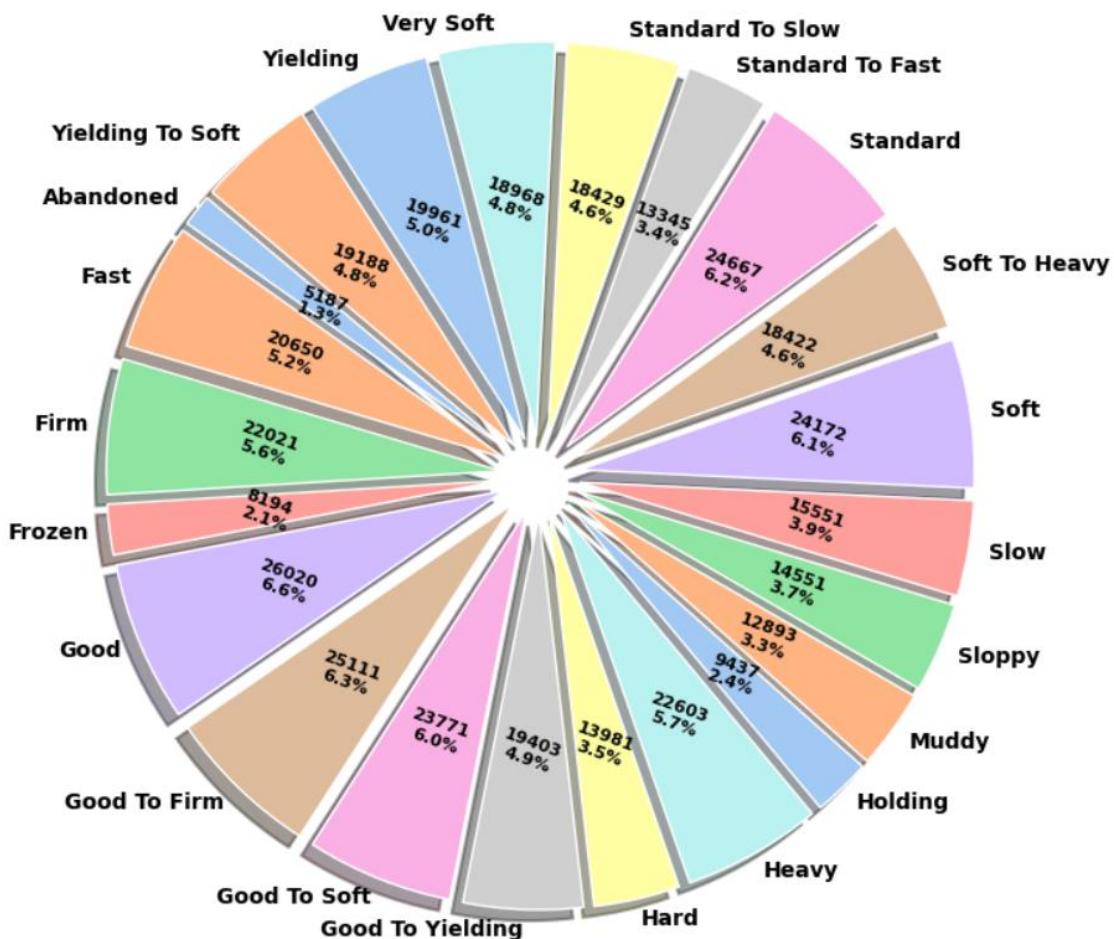


- התרשימים מדגים קשר ברור בין מצב המירוץ לזמן הנצחון, כאשר תנאי קרקע רכים וכבדים יותר מובילים לזמן מירוץ איטיים יותר. לעומת זאת, תנאי מזקקים ומהירים יותר מאפשרים זמני ניצחון מהירים יותר, מה שמראים כיצד משטחים שונים משפיעים על מצב המירוץ. תובנה זו יכולה להיות חיונית עבור מאמנים, מחרמים ומארגני מרוץ בעת חיזוי תוצאות מרוץ בתנאים משתנים.

### תוצאות ניתוח התפלגות תנאי המסלול לאורך השנים (סולם לוגריטמי):

התרשימים מציג את התפלגות של תנאי המסלול בכל השנים תוך שימוש בסולם לוגריטמי. הנתונים מראים בבירור כי תנאים כמו Standard, Good, Soft, ו-*Yielding* הם הנפוצים ביותר במהלך התקופה הנחקרת. באופן כללי, התנאים היבשים והרגילים יותר כמו *Good To Firm*, *Firm*, *Fast*, ו-*Yielding* היו בעלי שכיחות נמוכה יחסית. מצד שני, תנאים רטובים או כבדים יותר כמו *Heavy*, *Muddy*, ו-*Sloppy* היו גם הם פחות נפוצים, אך מופיעים בתדירות נמוכה יותר מתנאים יבשים.

## Distribution of Conditions Over All Years



- התרשימים מראה שתנאי מסלול כמו **Standard** ו-**Good, Soft** הם הנפוצים ביותר לאורך השנים, בעוד שתנאים יבשים ורטובים יותר היו פחות שכיחים.

### תוצאות ניתוח מטrixת הקורלציה:

מטrixת הקורלציה המוצגת מתרחשת את הקשרים בין המשתנים שנבחרו:

#### **Age, Position, WinningTime, Weight, DistanceKM**

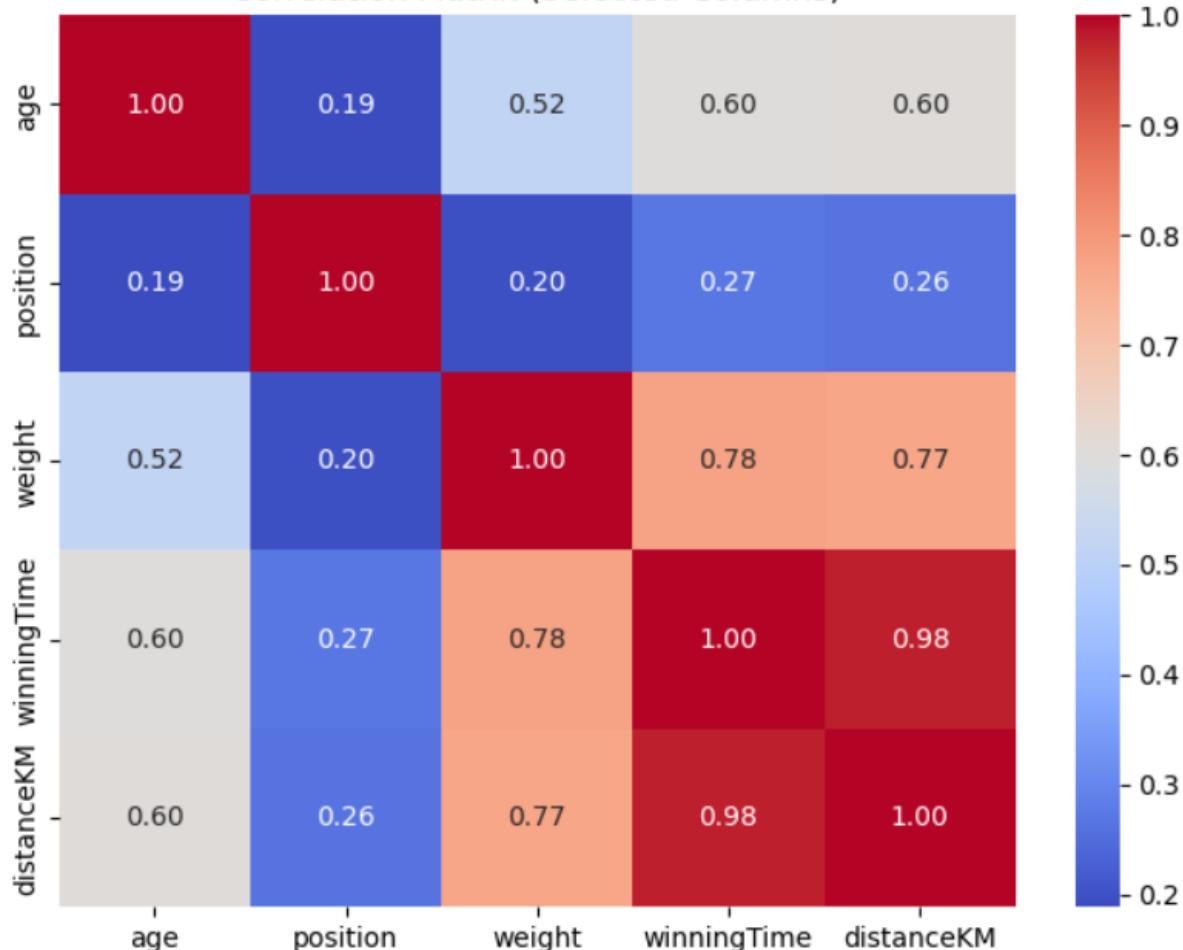
התוצאות מצביעות על כמה קשרים חזקים בין המשתנים:

קיים קשר חזק מאוד חיובי בין **WinningTime** ל - **DistanceKM** (0.98). ככלומר, ככל שהמרחק גדול יותר, זמן הזכיה נוטה להיות ארוך יותר, מה שיכל להשיע על תחרות קשה יותר במרוצים ארוכים.

קיים קורלציה חיובית בין **Weight** ל - **DistanceKM** (0.77) ול **WinningTime** (0.77). ככלומר, משקל גובה יותר קשור גם למרחק גדול יותר ולזמן זכייה ארוך יותר.

בין המשתנים **Weight** ו- **Age** ישנה קורלציה חיובית בינונית (0.52), מה שמעיד על כך שסוסים מבוגרים יותר נוטים לשקל יותר.

Correlation Matrix (Selected Columns)



- המטrixה מראה קשר חזק בין זמן זכייה למרחק המירוץ, וקורלציה חיובית בין משקל הסוס למרחק חמן זכייה, כאשר סוסים מבוגרים נוטים לשקל יותר.

## ניתוח עץ החלטות:

לאחר ביצוע הניתוח בעזרת עז החלטה לנביוי ניזחנות במרוצים, התקבלו המדים הבאים עבור סט הבדיקה:

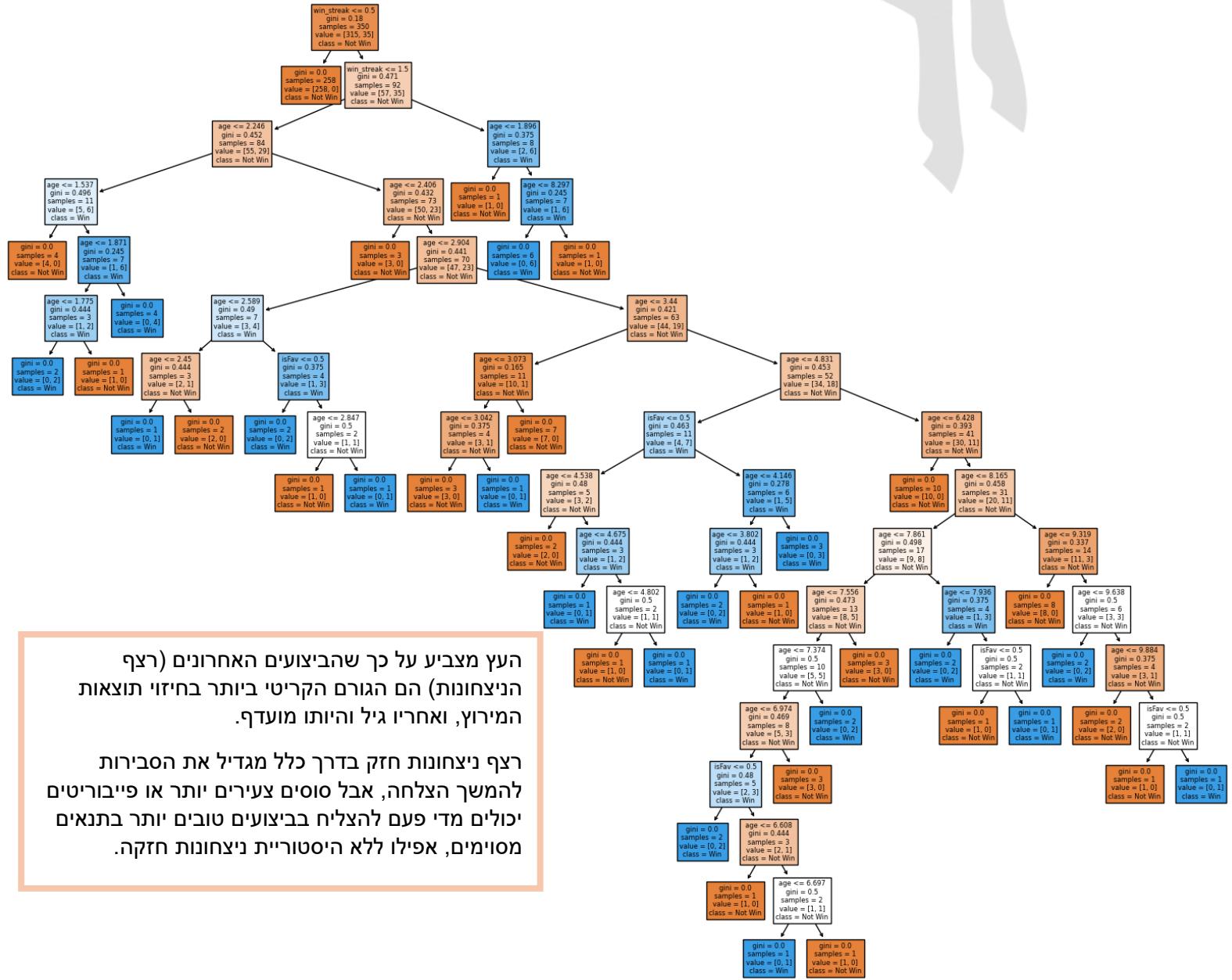
- דיוק:** (**Accuracy**) הממד מדד את היחס בין מספר התוצאות הנכונות לשאלה לבין מספר התוצאות הכוללות כל תוצאה נכונה ורעות. הערך שהתקבל הוא 0.73.

**דיוק התוצאות (Precision):** המודד מיצג את אחו התוצאות המדוקאות של הניצחונות מתור כלל התוצאות לניצחון, והערך שהתקבל הוא 0.72.

**ריגישות: (Recall)** הממד מייצג את אחוז הזיהוי הנכון של הניצחונות בפועל מתוך כלל הניצחונות, והערך שהתקבל הוא 0.71.

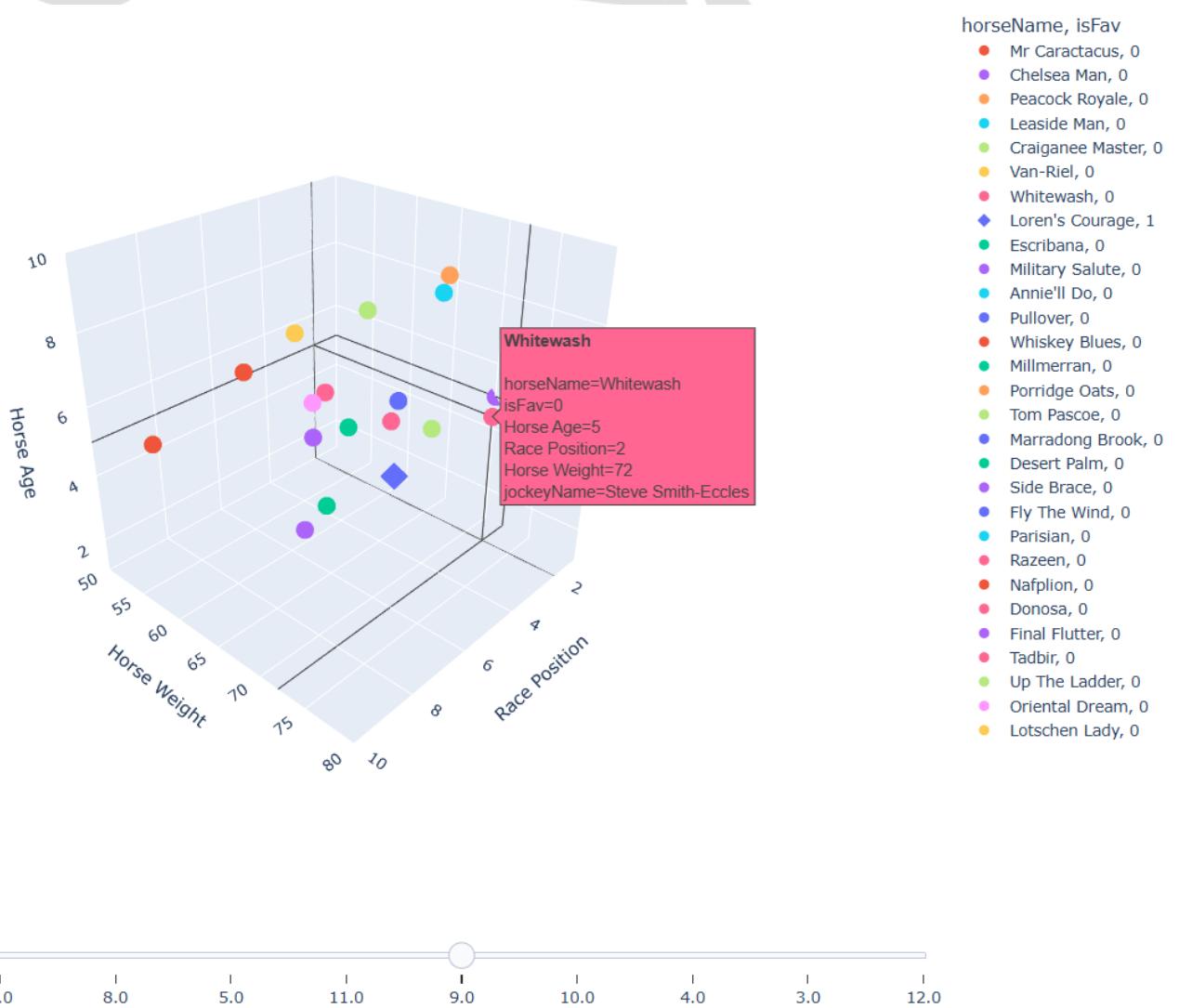
- מטריצת הבלבול (Confusion Matrix):** מטריצת הבלבול מדגימה את התפלגות התוצאות הנכונות והטעויות, תוך חלוקה לקטגוריות של ניצחון ופספוס בנייצחון.

## Decision Tree for Predicting Race Wins



### ניתוח גרפ' פיזור תלת-ממדי:

הקוד יוצר עליית פיזור תלת ממדית באמצעות **Plotly Express**, הממחיש את הקשר בין גיל הסוס, משקלו ומקום המירוץ. כל נקודה מייצגת סוס, עם מידע נוסף זמין ברכף.  
 כל פריט מייצג גיל סוס אחר. זה מאפשר להתבונן כיצד הקשר בין משקל ומקום הגזע משתנה על פני קבוצות גיל שונות.  
 כל סוס מיוצג על ידי צבע ייחודי, בנוסף יש סמל שבדיל בין סוסים אהובים (**isFav**) לבין סוסים שאינם אהובים ומופיעים פרטים נוספים כגון שם הרוכב.



- הגרף התלת-ממדי המראה את הקשר בין גיל, משקל, ומקום המירוץ, ומאפשר לבחון את ההשפעה של משתנים אלו על פני קבוצות גיל שונות.

# מסקנות:

## מסקנות על בסיס התפלגות גיל הסוסים במרוצים:

מהניתוח ניתן להסיק שהגיל המיטבי עבור סוסים להשתתף במרוצים הוא בין 3 ל-6 שנים. גילים אלו מייצגים את רוב הסוסים המתחרים, מה שמצויב על אף שבשלב זה של חיים, סוסים נמצאים בשיא כושרם להתחרות. סוסים מבוגרים יותר, במיוחד מעל גיל 10, פחות מיזגים, מה שעשו לעמידה על ירידה ביכולות הפיזיות או העדפה של בעלי הסוסים להפסיק את השתתפותם במרוצים בשלב זה.

## מסקנות לגבי ביצועי עשרה המאמנים המובילים:

מהניתוח ניתן להסיק כי ישנו מאמנים שצליחים להשיג אחוז ניצחונות גבוה למספר הסוסים שהם מאמנים קטנים יחסית.

הדבר יכול לעמידה על אסטרטגיות אימון יעילות יותר או על בחירה ממוקדת של תחרויות בהן יש סיכוי גבוהה יותר לניצחון. לעומת זאת, מאמנים אחרים משיגים אחוז ניצחונות נמוך יותר למקרים בהם מאמנים מסוימים מספר גדול יותר של סוסים, מה שיכול להוכיח על יכולות אימונם או על השתתפות בתחרויות חסרות סיכוי.

## מסקנות לגבי מגמות ביצוע של חמישה המאמנים המובילים לאורך השנים:

התוצאות במספר הניצחונות והמרוצים משקפות את האתגרים וההזהדרניות של מאמנים שונים מתמודדים איתם לאורך השנים. מאמנים שהצליחו לשמור על יציבות מסוימת הצלicho גם לשמר את מעמדם בזמן לאורך זמן.

ההצלחה של מאמן לאורך שנים תלואה בניהול נכון של משאבים והכנה מדויקת לארח המרוצים. המאמנים שהצליחו לשמור על יציבות יחסית לאורך השנים הם אלו שכנהרא נקבעו בגישה שיטתית ומושכלת.

## מסקנות לגבי אחוז הניצחונות של מאמנים מובילים בתנאי מירוץ שונים:

ניתן להסיק שמאנים מסוימים מתחכמים בתנאי מירוץ ספציפיים, מה שיכול להיות משמעותית בעת בחירת מאנון לסוס בתנאי מירוץ מסוימים.

בתנאי מירוץ קשים, רוב המאמנים אינם מצליחים לנצח, מה שיכול להוכיח על אף תנאים אלו קשים במיוחד עבור סוסים ומאמנים אחד.

### **مسקנות לגבי ביצועי סוסים פיבוריטים מול סוסים לא פיבוריטים:**

הנתונים יכולים לשמש לשיפור החלטות אסטרטגיות במרוצים, הן עבר מהמרים והן עברו מאמנים, בהתחשב בכך שסוסים פיבוריטים אכן מצלחים יותר, אך לא מבטחים ניצחון מוחלט.

### **مسקנות מהשפעת הגיל על הסתרות לניצחון במרוצים:**

מהגרף ניתן להסיק שגיל הסוס הוא גורם משמעותי בקביעת ההסתבות לניצחון במרוץ. סוסים צעירים יותר, בעיקר בגילאים 2-5, מראים הסתרות גבוהה יותר לניצחון, בעוד שסוסים מבוגרים יותר מראים ירידה בהסתבות זו. המסקנה היא שיש לשים דגש על גיל הסוס כאשר מנתחים את סיכויו לניצחון במרוץ. יתכן שהוא יכול להיות כלי עזר בקבלת החלטות אסטרטגיות בעת בחירת סוסים פיבוריטים במרוצים.

### **مسקנות לגבי השפעת מצב המסלול על זמן הזכהיה:**

ניתן להסיק שמצוותי מסלול יותר 'כבדים' או 'רטובים' כמו **Soft To Heavy** ו-**Heavy** דורשים מהסוסים יותר זמן לסיים את המירוץ. בעוד שמצוותי מסלול 'קלים' יותר כמו **Fast** מאפשרים זמנים מהירים יותר. תוצאה זו עשויה לשקוף את הקושי הפיזי הנדרש מהסוסים כאשר הם מתמודדים עם תנאי מסלול פחות אידיאליים. ההבדלים בזמןי הזכהיה בין מצווותי המסלול השונים עשויים להיות קרייטיים עבור מאמנים, בעלי סוסים, ומהרמים, שכן הם מצבעים על כך שהתנאים המסלוליים יכולים להשפיע משמעותית על ביצועי הסוסים.

### **مسקנות לגבי השפעת התפלגות תנאי המסלול לאורך השנים:**

מהניתוח ניתן להסיק כי תנאי המסלול המשפיעים ביותר על תוצאות המרוצים היו אלה שהופיעו בתדרות גבוהה יותר כמו **Soft Good** ו-**Good**. המשמעות היא שה坦אים האלה כנראה היו סטנדרטיים והעדיפים עבור מרבית המרוצים, מה שمعد על יציבות מסוימת בתנאי המסלול לאורך השנים. שימוש בסולם לוגריטמי להדגשת תנאים שהיו פחות נפוצים אך עדין חשובים, כמו **Abandoned**, **Fast**, **Frozen** תנאים נדירים במרוצי סוסים.

### **مسקנות לגבי קשרי הקורלציה בין המשתנים:**

הניתוח מצבע על כך שמרקח המירוץ הזמן הזכהיה הם משתנים קשורים באופן ישיר וחזק זה לזה, מה שמדגיש את השפעת המרחק על תוצאות המירוץ. בנוסף, הקורלציה החיובית בין משקל הסוס לבין המרחק הזמן הזכהיה מרגמת משקל רב עשוי להיות גורם מכני במרוצים ארוכים. הקשר בין הגיל והמשקל מצבע על כך שסוסים מבוגרים נוטים להיות כבדים יותר, מה שעשי להשפיע על הביצועים שלהם בתנאים מסוימים.

### **مسקנות ניתוח עץ החלטות:**

העץ מורכב למדים עם פיצולים רבים, מה שמצוין על כך שהזיווי תוצאות להצלחת הסוס מושפע ממספר גורמים. צמתים מסוימים בעץ מובילים לתחזיות בעלות ביטחון גבוה (צמתים כתומים או כחולים), בעוד אחרים יש תוצאות מעורבות יותר. ניתן להשתמש בעץ ההחלטה זהה כדי לנתח ולהזות את תוצאות המירוץ על סמך נתוני עבר של הסוס, ולהיעזר ולחזות גורמי מפתח שתורמים לניצחון.

### **מסקנות עבור גרפ' פיזור תלת-ממדי:**

ההדמייה מספקת מבט מקיף על משחק הגומלין המורכב בין גיל הסוס, משקלו וביצועי המירוץ.

זה מדגיש את האופי הרב-גוני של מרכיבי סוסים, שבו גורמים רבים תורמים להצלחת הסוס מעבר לתכונות פיזיות כמו גיל ומשקל.

