

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

גל שמואל  
רחל יהואלשט



# מבוא

מרוצי סוסים הם אחת מהפעילויות הספורטיביות העתיקות ביותר בעולם, והם שומרים על מעמד הפופולרי גם בעידן המודרני. ספורט זה, שמתחיל עוד מתקופת רומא העתיקה, עבר שינויים והתפתחויות רבות לאורך ההיסטוריה, אך נשאר נאמן למורשתו, כשהוא משלב בין כישרון הרכיבה, אימון סוסים ברמה גבוהה וידע מעמיק על התנהגותם הפיזיולוגית והפסיכולוגית של בעלי חיים אלו.

מעבר להיותם ספורטיבי, מרוצי הסוסים מהווים מרכיב משמעותי בתרבויות רבות, בעיקר במדינות כמו בריטניה, ארצות הברית, אוסטרליה ויפן. בכל אחת ממדינות אלו מתקיימים אירועי מרוצי סוסים יוקרתיים, המושכים אליהם קהל רב ומציבים אתגרים ייחודיים הן לסוסים והן לרוכבים.

בנוסף, תעשיית מרוצי הסוסים מהווה חלק בלתי נפרד מהכלכלה העולמית, כאשר היא כוללת ענפים רבים כמו הימורים, מסחר בחיות, גידול סוסים גזעיים, תעשיית הווטרינריה והאימון, ותרבות עשירה המתפתחת סביב הפעילות הזו. השוק של מרוצי הסוסים הוא שוק דינמי ובעל נפח כלכלי עצום, כאשר ההימורים לבדם 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 משתמשים במגוון שיטות מתקדמות לשיפור הדיוק בתחזיות ולהגדלת הרווחיות.

במסגרת פרויקט זה, אנו מתמקדים בניתוח נתונים היסטוריים של מרוצי סוסים משנים 1990-2020. המטרה היא לבחון את ההשפעה של שיטות ניתוח מתקדמות על תחום מרוצי הסוסים ולבדוק כיצד ניתן להשתמש במידע זה כדי ליעל את תחזיות התוצאות ואת קבלת ההחלטות בתחום זה. ניתוח זה יכלול בחינת תכונות ומסלולים, ויתמקד במציאת קשרים ומגמות שיכולים לסייע שונות כמו נתוני סוסים בניבוי תוצאות המרוצים בצורה מדויקת יותר.

# הבעיה

במרוצי סוסים, כמו בספורט מקצועי אחר, ההצלחה תלויה במספר רב של משתנים, כגון בריאות הסוס, יכולות הרוכב, תנאי המסלול. עם זאת, למרות השפע של הנתונים הזמינים, רבים מההימורים ומקבלי ההחלטות בתעשייה זו מסתמכים על אינטואיציה וניסיון אישי, מה שמוביל לרמת חיזוי נמוכה ולסיכון כלכלי גבוה.

האתגר המרכזי בפרויקט זה הוא למצוא דרך לשפר את היכולת לחזות תוצאות מרוצי סוסים בצורה מדויקת יותר על בסיס ניתוח נתונים. הפער הקיים כיום הוא חוסר במודלים חיזוי מתקדמים שמסוגלים לשקלל את כל המשתנים הרלוונטיים ולהציע תחזיות אמינות. הפער הזה לא רק פוגע ברווחיות של המשתתפים בתעשיית ההימורים, אלא גם מונע שימוש אופטימלי במידע הזמין לצורך שיפור ביצועים והקטנת 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 המרוצים ולספק תחזיות מדויקות יותר שיכולות להוביל לקבלת החלטות מושכלת יותר, הן עבור המתחרים והן עבור ההימורים.

# השיטה

בהתאם לאתגרים בניתוח תוצאות מרוצי סוסים, נעשה שימוש בשיטות ניתוח מתקדמות ובמגוון כלים וספריות בפייתון. העבודה החלה בהכנת הנתונים, המשיכה בניתוח ראשוני, והתקדמה לניתוח מעמיק.  
הכנת הנתונים:

בשלב הראשון של העבודה, ביצענו אחזור והכנה של נתוני המרוצים ממאגרי מידע חיצוניים. השתמשנו ב-Kaggle API להורדת מערכי נתונים המכילים מידע על סוסים ומרוצים משנת 1990 עד שנת 2020, נעזרנו ב ספריות os ו-json לטעינה ובספריית zipfile כדי לחלץ את תוכנו לתוך התיקיה המיועדת.

לאחר מכן השתמשנו בקוד לאיחוד נתוני 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 AI	7.0	0	Buddy Lammas	Ryan Bishop	11	0.016393	12.0	2.75	57

וקוד לאיחוד נתוני המרוצים (races) ל-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 horses\_data ו-races\_data ל-DataFrame בשם **merged\_data**.

	rid	horseName	age	isFav	trainerName	jockeyName	position	decimalPrice	saddle	position1	...	time	title	ages	date	distance	countryCode	condition	hurdles	winningTime	distanceKM
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	Mandraki 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	Turnberry 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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1107310	415090	Beefeater	8.0	0	Roydon Bergerson	Hazel Schofer	7	0.030303	6.0	shd	...	03:45	Rydgcs Wellington Captain Cook Stakes (Group 1...	2yo+	2020	1m	NZ	Soft	NaN	99.66	1.609340
1107311	415090	Aimee's Jewel	4.0	0	Trudy Keegan	Lisa Allpress	8	0.153846	11.0	.5	...	03:45	Rydgcs Wellington Captain Cook Stakes (Group 1...	2yo+	2020	1m	NZ	Soft	NaN	99.66	1.609340
1107312	415090	Times Ticking	5.0	0	Alby Macgregor	Jonathan Riddell	9	0.044053	8.0	hd	...	03:45	Rydgcs Wellington Captain Cook Stakes (Group 1...	2yo+	2020	1m	NZ	Soft	NaN	99.66	1.609340
1107313	415090	Shadows Cast	8.0	0	Mark Oulaghan	Johnathan Parkes	10	0.041152	3.0	1	...	03:45	Rydgcs Wellington Captain Cook Stakes (Group 1...	2yo+	2020	1m	NZ	Soft	NaN	99.66	1.609340
1107314	415090	Awesome Al	7.0	0	Buddy Lammis	Ryan Bishop	11	0.016393	12.0	2.75	...	03:45	Rydgcs Wellington Captain Cook Stakes (Group 1...	2yo+	2020	1m	NZ	Soft	NaN	99.66	1.609340

## שיטת העבודה וניתוח ראשוני:

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

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

- **Numpy**: שימשה אותנו לביצוע חישובים מתמטיים כגון חישוב ממוצעים, סטיות תקן ומניפולציות על מערכים.
- **Seaborn**: אפשרה יצירת גרפים ויזואליים כמו **heatmaps** לזיהוי דפוסים ומטריצות קורלציה.
- **Matplotlib** ו-**Plotly**: להצגת קשרים בין משתנים מרכזיים ולחקירת הנתונים
- באופן אינטראקטיבי כמו יצירת היסטוגרמות, דיאגרמות פיזור ו-**line charts**.
- **Sklearn**: שימשה לפיצול הנתונים ולהכשרת מודלים כמו רגרסיה לינארית ועץ החלטה, והערכנו את ביצועי המודלים באמצעות מדדים שונים.



## ניתוח מעמיק:

בשלב הניתוח המעמיק, עברנו להערכה של קשרים מורכבים יותר בין משתנים שונים באמצעות טכניקות מתקדמות.

בנינו מודלים של רגרסיה 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 בעץ החלטה לצורך ניתוח הסתברותי של ניצחונות הסוסים. המודל אפשר לנו לזהות את הגורמים המשפיעים ביותר על הסיכוי לניצחון, כגון גיל הסוס ותנאי המסלול. ראינו שקיים קשר חזק בין תנאי המסלול לבין סיכוי לניצחון, כאשר תנאים טובים יותר העלו משמעותית את הסיכויים לניצחון.

כדי להמחיש את התוצאות ולחקור את הנתונים באופן אינטראקטיבי, השתמשנו בספריית **Plotly** ליצירת גרפים תלת-ממדיים שהתמקדו בקשרים בין גיל הסוס, משקלו והמיקום שבו הוא הגיע. ההדמיות האינטראקטיביות הללו אפשרו לנו להבין בצורה ויזואלית את התפלגות הנתונים ואת הקשרים המרכזיים בין המשתנים.



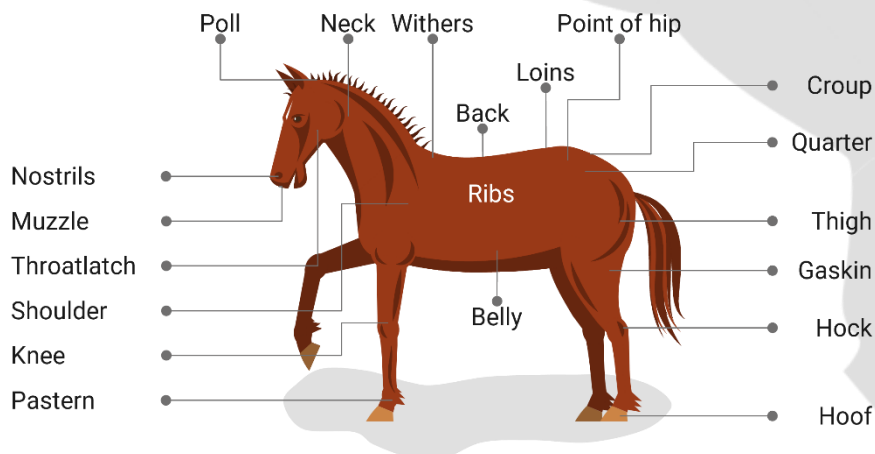
# תיאור הדאטאסטים

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, לפרויקט שלנו התמקדנו בעמודות שנראו לנו הכי רלוונטיות כמו:

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

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



הקבצים של races מכילים נתונים שנערכו בין השנים 1990-2020, הנתונים על המרוצים עצמם, מהדאטאסט המצויין לקחנו את העמודות העיקריות הבאות:

- **Rid**: מזהה המירוץ.
- **Course**: מסלול
- **Time**
- **Title**: שם המירוץ
- **Ages**: מאיזה גיל אפשר להתחרות במסלול
- **Distance**: אורך המסלול.
- **CountryCode**: איפה המסלול התקיים
- **Condition**: תנאי המירוץ.
- **Hurdles**: כמות המכשולים שיהיו במסלול.
- **WinningTime**: משך הזמן שלקח לסוס במקום הראשון לנצח

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

בנוסף יצרנו **DataFrame** בשם **merged\_data** 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 תהליך המחקר:

תהליך המחקר שלנו התבסס על גישה רב-שלבית, שכללה את השלבים הבאים:

## 1. איסוף של הנתונים:

התחלנו את הפרויקט באיסוף נתונים ממאגרי מידע חיצוניים עבור מרוצי סוסים מהשנים 1990-2020, תוך שימוש ב-Kaggle API להורדה של Dataset שנקרא **horse-racing** ובו קבצים המכילים מידע על הסוסים ועל המרוצים. בעזרת הספריות **os** ו-**json**, טענו את פרטי ההתחברות ל-Kaggle מתוך קובץ **JSON** ששמר את שם המשתמש והמפתח (**API key**). השתמשנו בפרטים אלו כדי להגדיר את משתני הסביבה הנדרשים לביצוע בקשות ל-Kaggle API. לאחר מכן, השתמשנו ב-Kaggle API להורדת ה-Dataset של מרוצי הסוסים, ועשינו שימוש בספריית **zipfile** כדי לחלץ את תוכנו לתוך התיקיה המיועדת. ווידאנו את ההצלחה של החילוץ 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 עבור כל קבוצת קבצים כלל טעינה של קבצי ה-CSV עבור כל שנה בנפרד, מיצוי העמודות הרלוונטיות, והוספת 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 המרת יחידות מדידה (לדוגמה, ממיל לקילומטר), זיהוי ערכים חסרים וטיפול בהם, ושמירה על עמודות חשובות בלבד לצורך ניתוח.

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

### 3. ניתוח ראשוני

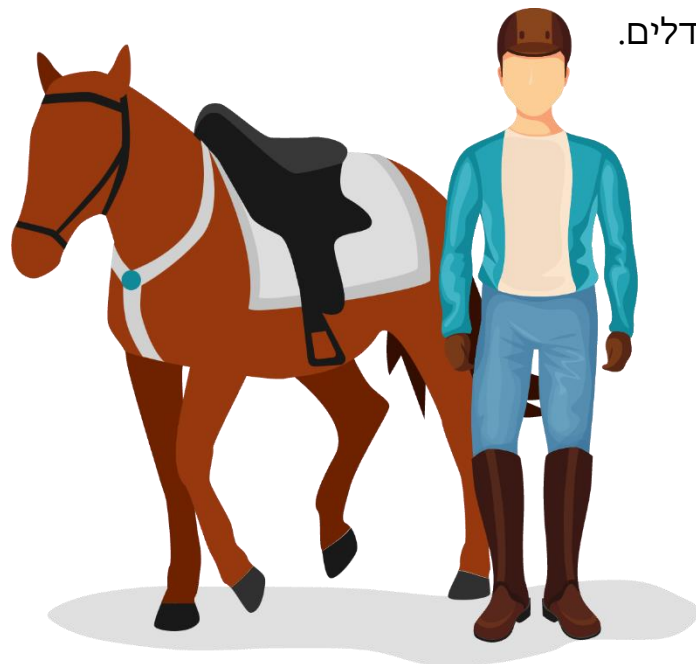
בשלב הניתוח הראשוני, השתמשנו בכלים כמו **Pandas** ו-**Seaborn** לחקירת המגמות הבסיסיות בדאטה. חקרנו את הקשרים הבסיסיים בין משתנים כמו גיל הסוס והמשקל שלו לבין המיקום שבו סיים את המירוץ. יצרנו טבלאות ציר וגרפים ויזואליים (כמו **scatter\_3d Histogram**) על מנת לקבל תמונה ראשונית של ההתפלגויות והקורלציות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

### 4. בניית מודלים לחיזוי

בשלב הבא, התמקדנו בבניית מודלים לחיזוי תוצאות המרוצים. השתמשנו במודלים שונים כמו **Linear regression** ו-**Decision Trees**. כל מודל עבר תהליך של אימון, בדיקה והערכה, 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

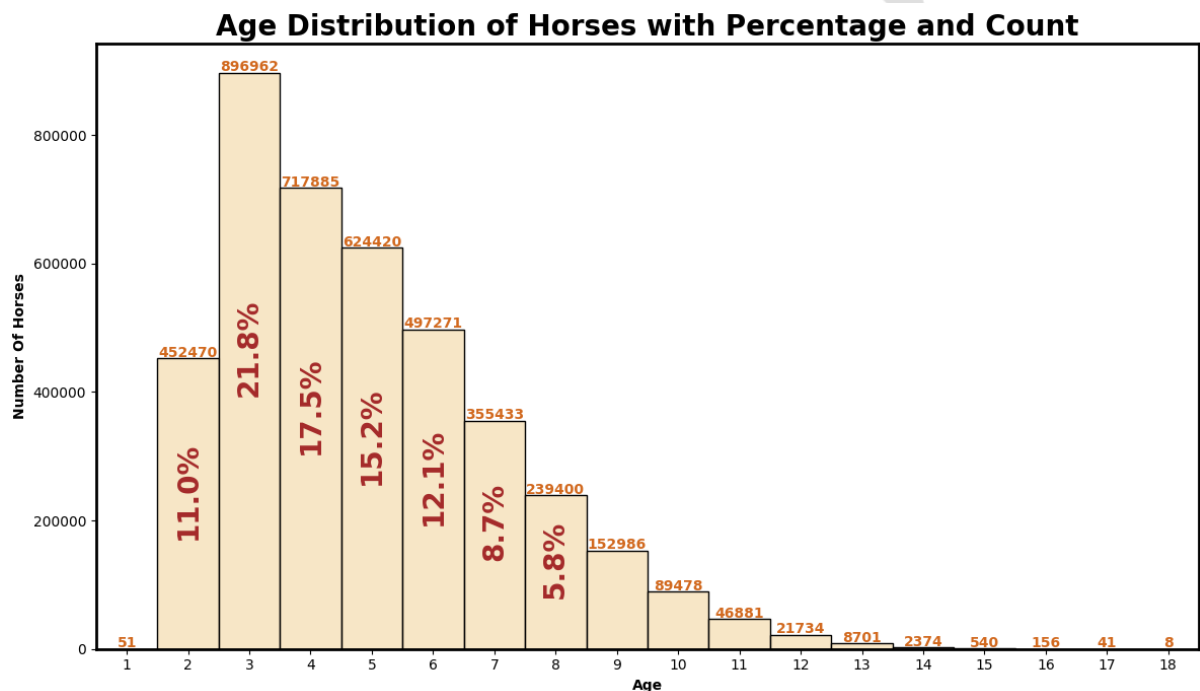
לאחר בניית המודלים, השתמשנו בספריות הדמיה כמו **Plotly** ו-**sklearn** ליצירת גרפים אינטראקטיביים שהתמקדו בקשרים המורכבים בין משתנים. הדמיות אלו עזרו לנו להבין באופן ויזואלי את הקשרים המרכזיים בין המשתנים ולגלות דפוסים נסתרים בדאטה. תהליך המחקר לווה לאורך כל הדרך בבדיקות תיקוף והערכה, כאשר נבדקו ביצועי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



# תוצאות:

## ניתוח התפלגות גיל הסוסים והשפעתו על השתתפות במרוצים:

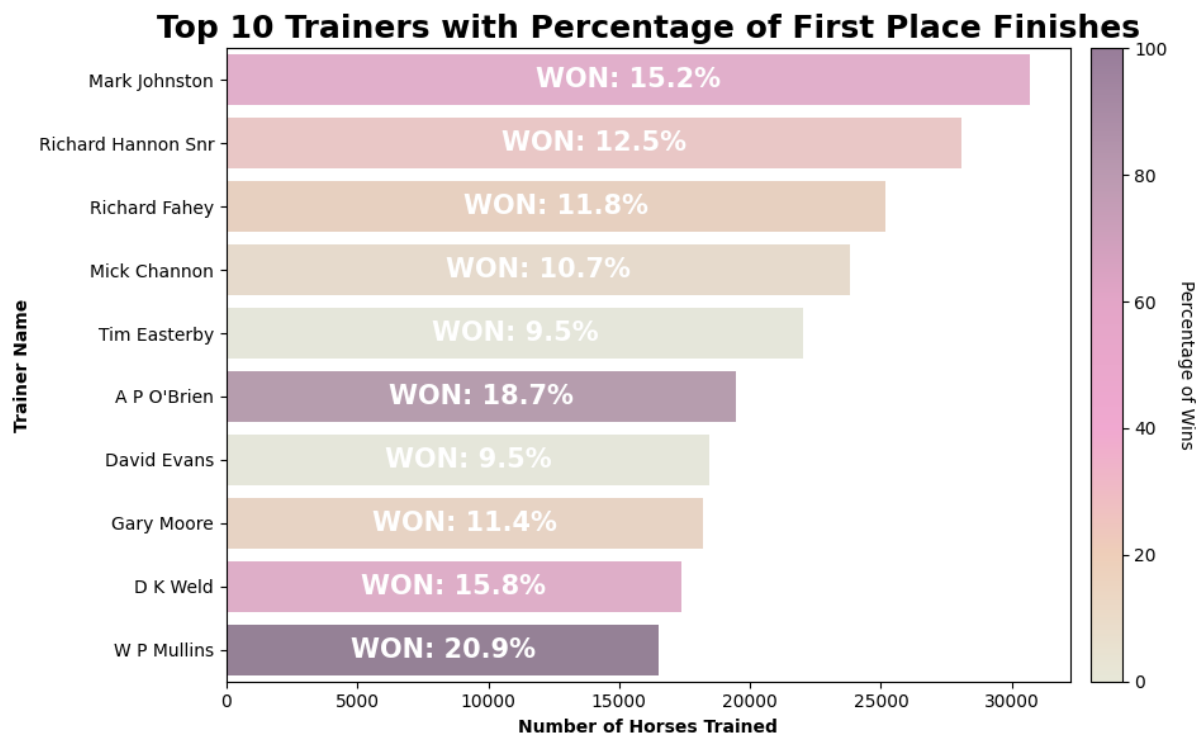
הגרף מתאר את התפלגות הגילאים של סוסים שהשתתפו במרוצים לאורך השנים. מהגרף ניתן לראות כי רוב הסוסים המתחרים הם בגילאי 3 עד 6 שנים. גיל 3 הוא הנפוץ ביותר, עם 896,962 סוסים שהם כ-21.8% מכלל הסוסים במאגר הנתונים. הגילאים 4 ו-5 גם הם מייצגים אחוזים משמעותיים, עם 17.5% ו-15.2% בהתאמה. לאחר מכן, חלה ירידה מתמדת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



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

### ניתוח אחוזי ניצחונות של עשרת המאמנים המובילים:

הגרף מציג את עשרת המאמנים המובילים במספר הסוסים שהם אימנו, יחד עם אחוזי הניצחונות שלהם. מהנתונים ניתן לראות כי המאמן עם אחוז הניצחונות הגבוה ביותר הוא **W P Mullins** עם 20.9% ניצחונות, למרות שמספר הסוסים שהוא אימן נמוך יחסית לאחרים. לעומתו, **Mark Johnston**, המאמן שאימן את המספר הגבוה ביותר של סוסים (מעל 30,000), הגיע רק לאחוז ניצחונות של 15.2%.



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

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

### מגמות כלליות לאורך השנים:

המגמות העיקריות שנראות בגרף מצביעות על כך שכל אחד מחמשת המאמנים המובילים עבר תנודות משמעותיות במספר הניצחונות והמרוצים לאורך השנים. ניתן לראות שמאמן מסוים עשוי לבלוט במיוחד באחת השנים, ואז לאבד את יתרונו בשנים אחרות.

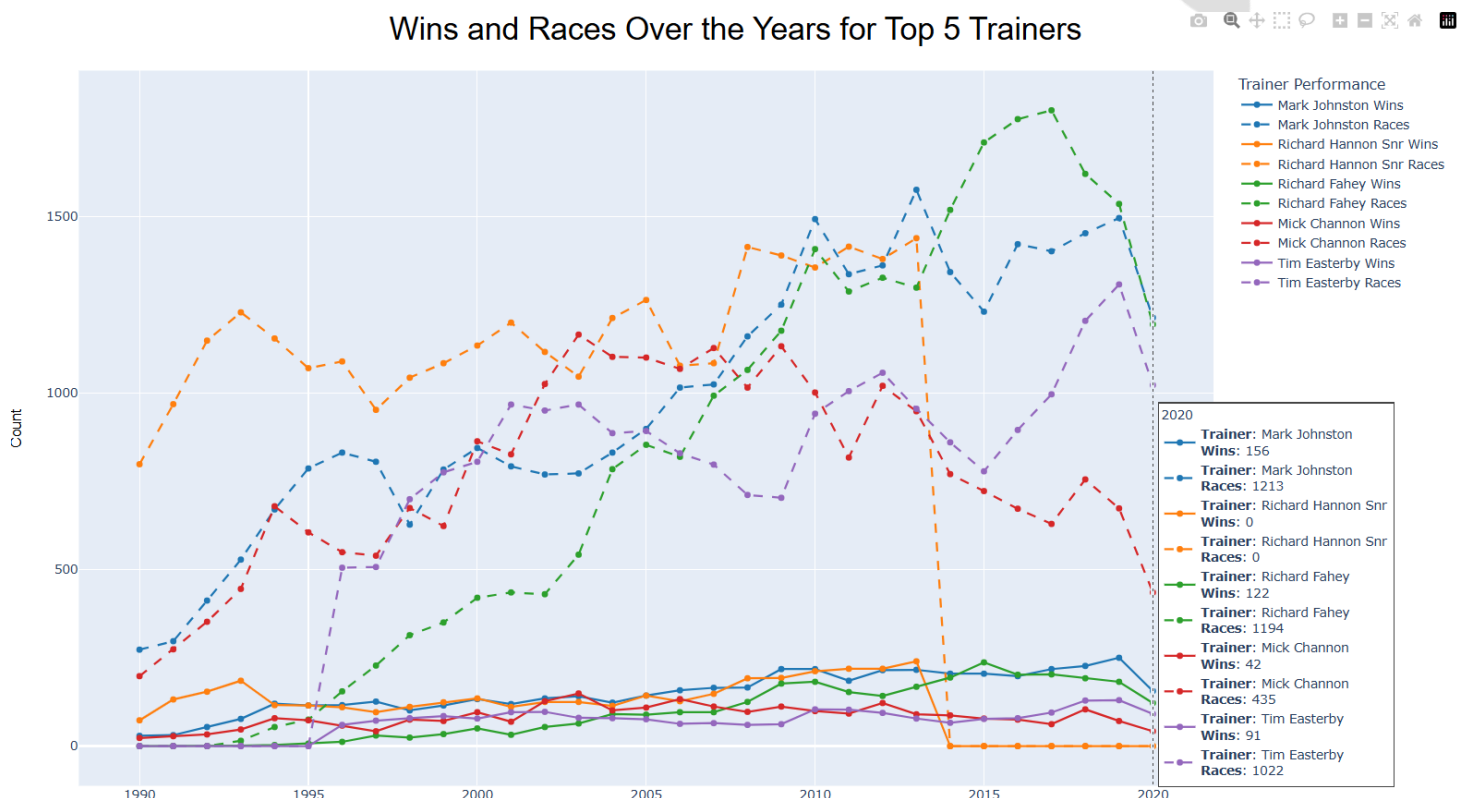
**Richard Hannon Snr**: מציג מגמת עלייה בולטת בניצחונות עד לשיא בסביבות 2013, אך לאחר מכן יש ירידה חדה במספר הניצחונות והמרוצים.

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

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

- ניתן לראות התאמה כללית בין מספר המרוצים שמאמן מסוים משתתף בהם לבין מספר הניצחונות שלו, אך לא תמיד מדובר בהתאמה מדויקת. לדוגמה, בשנים מסוימות ישנם מרוצים רבים אך מספר הניצחונות נמוך יחסית.

Wins and Races Over the Years for Top 5 Trainers

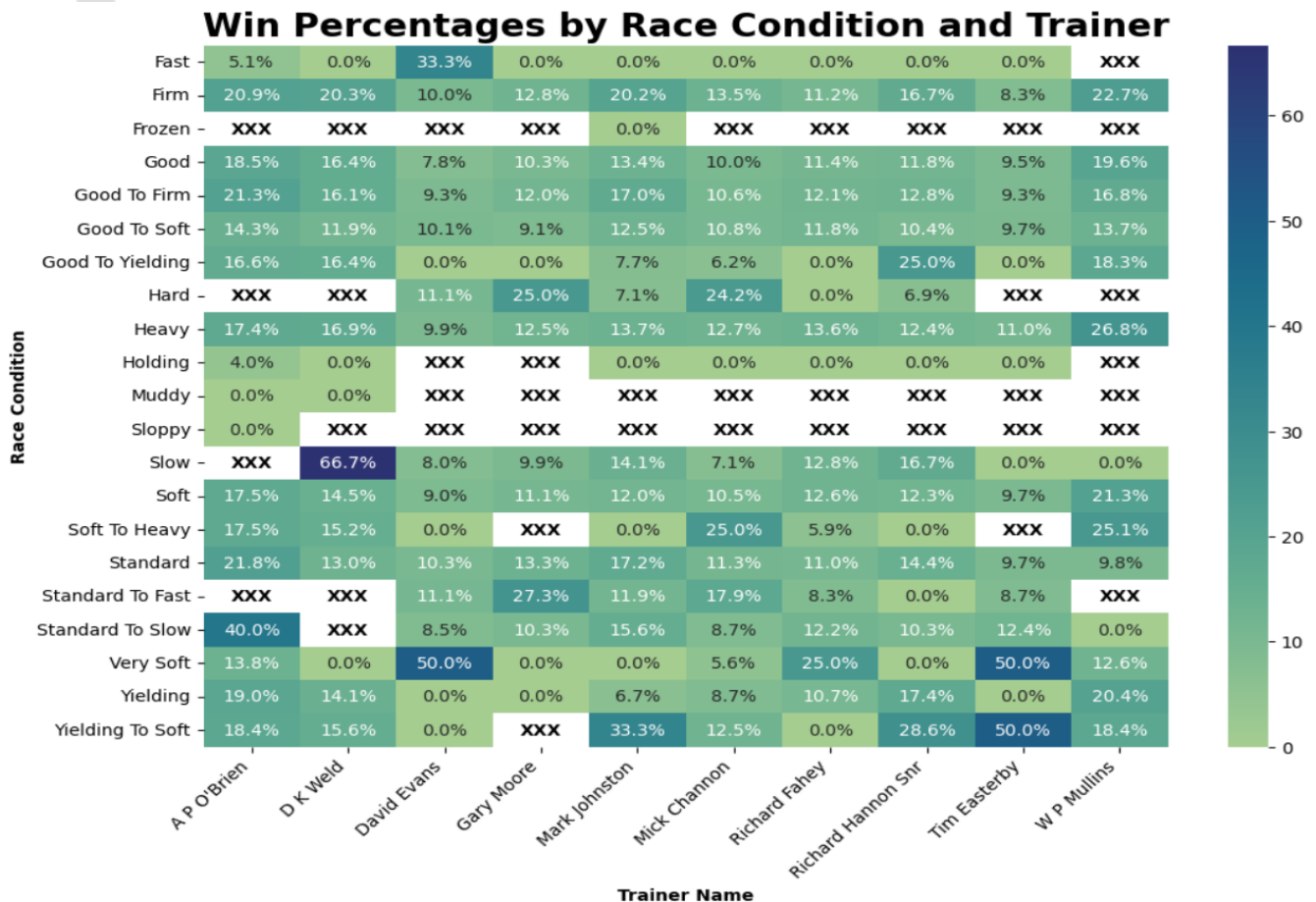


## ניתוח אחוזי הניצחונות של מאמנים מובילים בתנאי מירוץ שונים:

אחוזי ניצחונות בתנאי מירוץ איטיים (**Slow**):  
המאמן "**Gary Moore**" השיג את אחוז הניצחונות הגבוה ביותר (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.  
אחוזי ניצחונות בתנאי מירוץ סטנדרטיים (**Standard To Slow**):  
המאמן "**Tim Easterby**" השיג 50% אחוזי ניצחון בתנאי מירוץ סטנדרטיים איטיים.

שונות באחוזי ניצחונות של מאמנים אחרים:  
מאמנים אחרים מציגים אחוזי ניצחונות שונים בתנאים שונים, עם תאים רבים המראים את הערך "**XXX**" בשל חוסר נתונים או השתתפות באותם תנאי מירוץ.



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

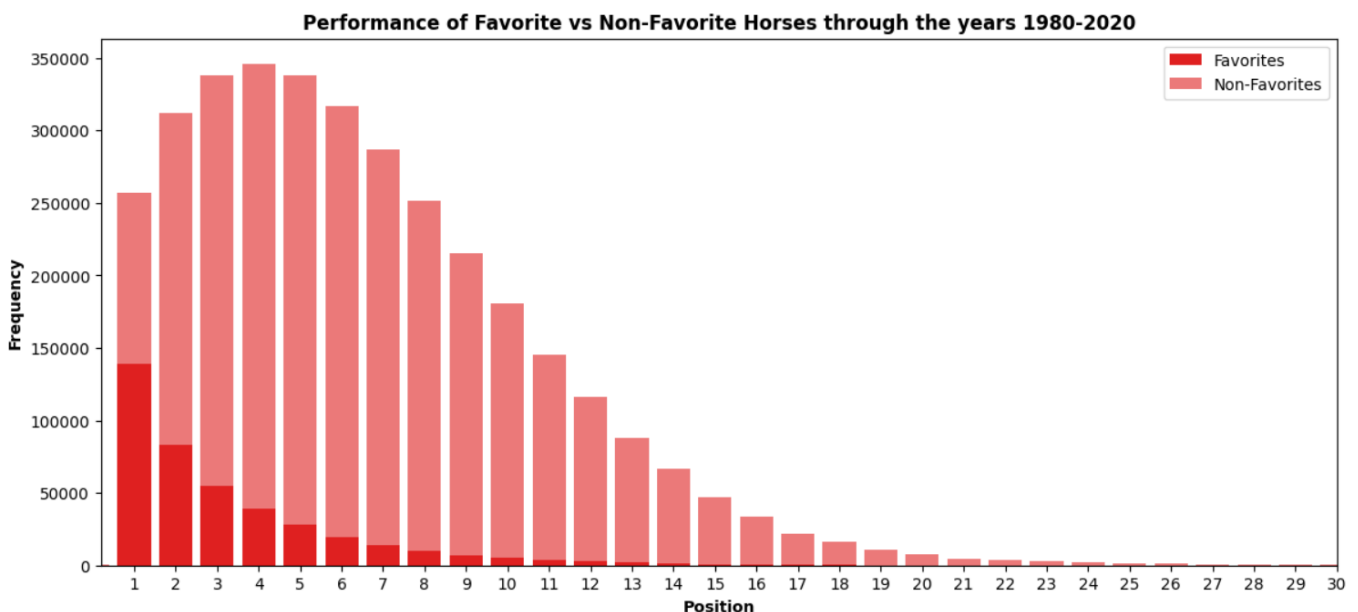
## ניתוח ביצועי סוסים פייבוריטים מול סוסים לא פייבוריטים:

### סוסים פייבוריטים:

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

### סוסים לא פייבוריטים:

הסוסים הלא פייבוריטים (**Non-Favorites**) מופיעים בכמויות גדולות יותר במיקומים 4 ומטה, עם שכיחות גבוהה במיוחד במיקומים 5-10. במיקום הראשון, הסוסים הלא פייבוריטים מופיעים בשכיחות נמוכה יחסית.

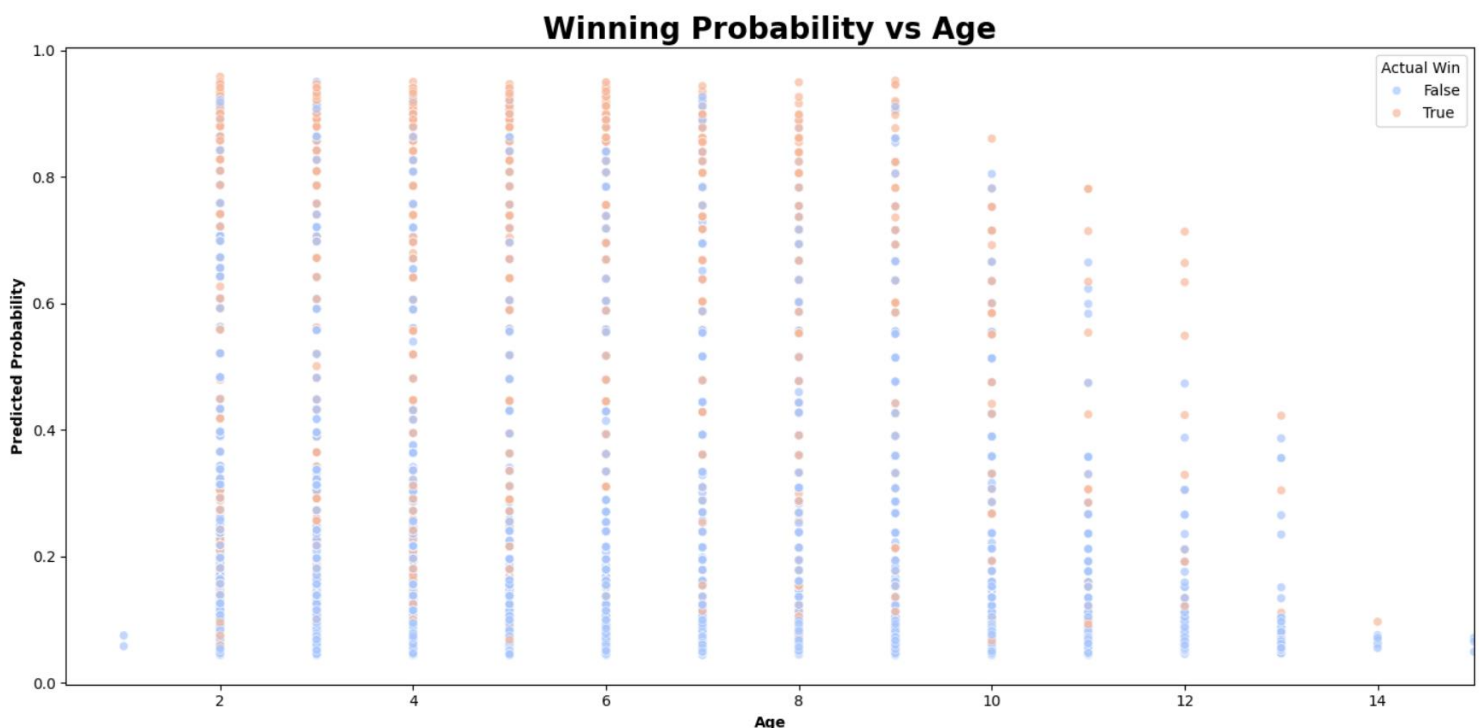


- הגרף מראה שהפייבוריטים מצליחים יותר ומגיעים לעיתים קרובות יותר למקומות גבוהים, בעוד שסוסים לא פייבוריטים ממוקמים בעיקר במקומות הנמוכים.

## ניתוח הסתברות לניצחון של סוסים לפי גיל: (תוצאות מודל רגרסיה לוגיסטית)

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

נראה כי הסוסים בגילאי 2-5 הם בעלי ההסתברות הגבוהה ביותר לניצחון, כפי שנראה מההתקבצות של נקודות עם הסתברות גבוהה באזורים אלו. מעבר לגיל 8, ההסתברות לניצחון הולכת ויורדת, 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 ירידה בהסתברות לניצחון עם העלייה בגיל.



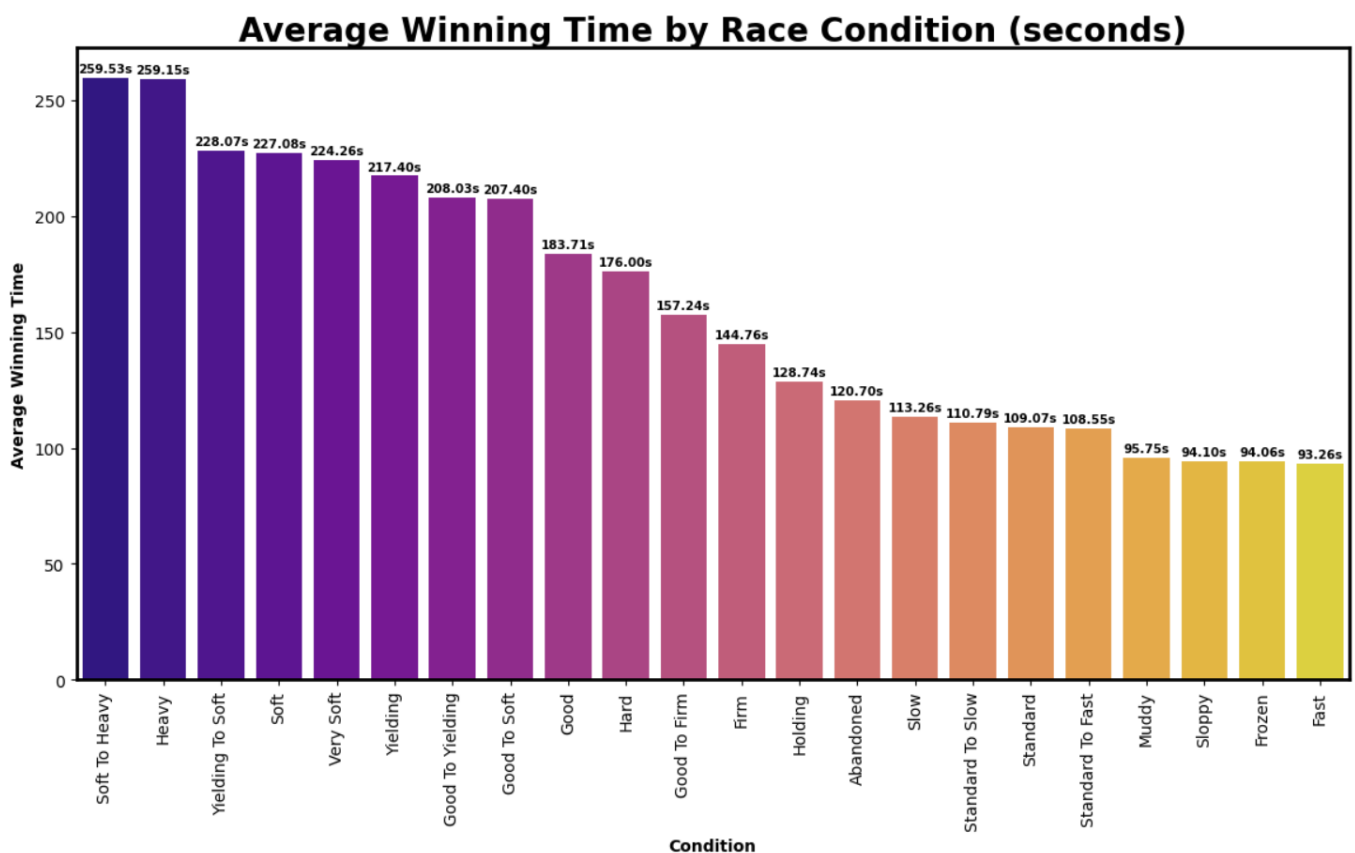
- הגרף מראה שההסתברות לניצחון היא הגבוהה ביותר עבור סוסים בגילאי 2-5 ויורדת באופן כללי עם העלייה בגיל, במיוחד לאחר גיל 8.



### תוצאות ניתוח זמן זכייה ממוצע לפי מצב מסלול:

הגרף המוצג מציג את זמן הזכייה הממוצע לפי מצב מסלול במרוצי סוסים, בו ערכי הזמן נמדדים בשניות. הזמנים הממוצעים חושבו לכל אחד ממצבי המסלול הקיימים בנתונים, ולאחר מכן הגרף ממוין בסדר יורד.

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

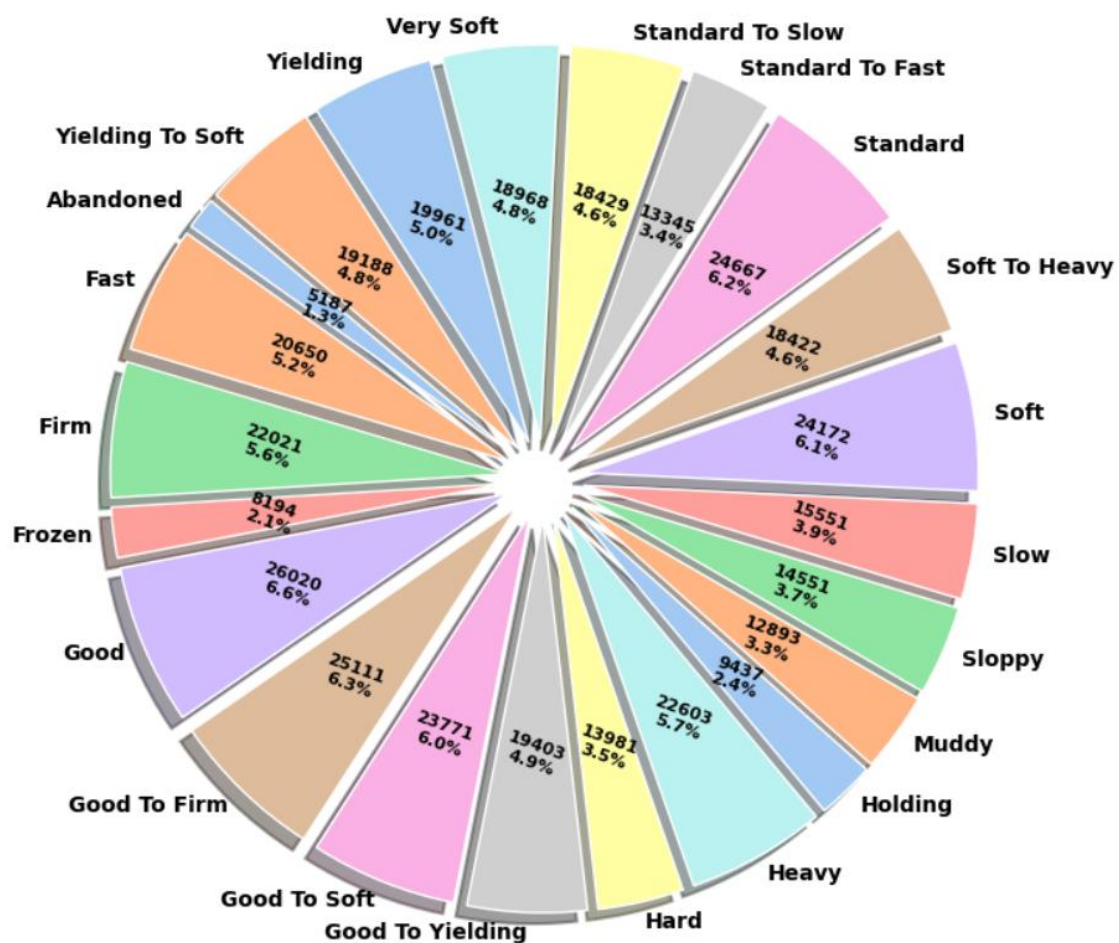


- התרשים מדגים קשר ברור בין מצב המירוץ לזמני הניצחון, כאשר תנאי קרקע רכים וכבדים יותר מובילים לזמני מירוץ איטיים יותר. לעומת זאת, תנאים מוצקים ומהירים יותר מאפשרים זמני ניצחון מהירים יותר, מה שמראים כיצד משטחים שונים משפיעים על קצב המירוץ. תובנה זו יכולה להיות חיונית עבור מאמנים, מהמרים ומארגני מירוץ בעת חיזוי תוצאות מירוץ בתנאים משתנים.

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

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

### Distribution of Conditions Over All Years



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

### תוצאות ניתוח מטריצת הקורלציה:

מטריצת הקורלציה המוצגת מתארת את הקשרים בין המשתנים שנבחרו:

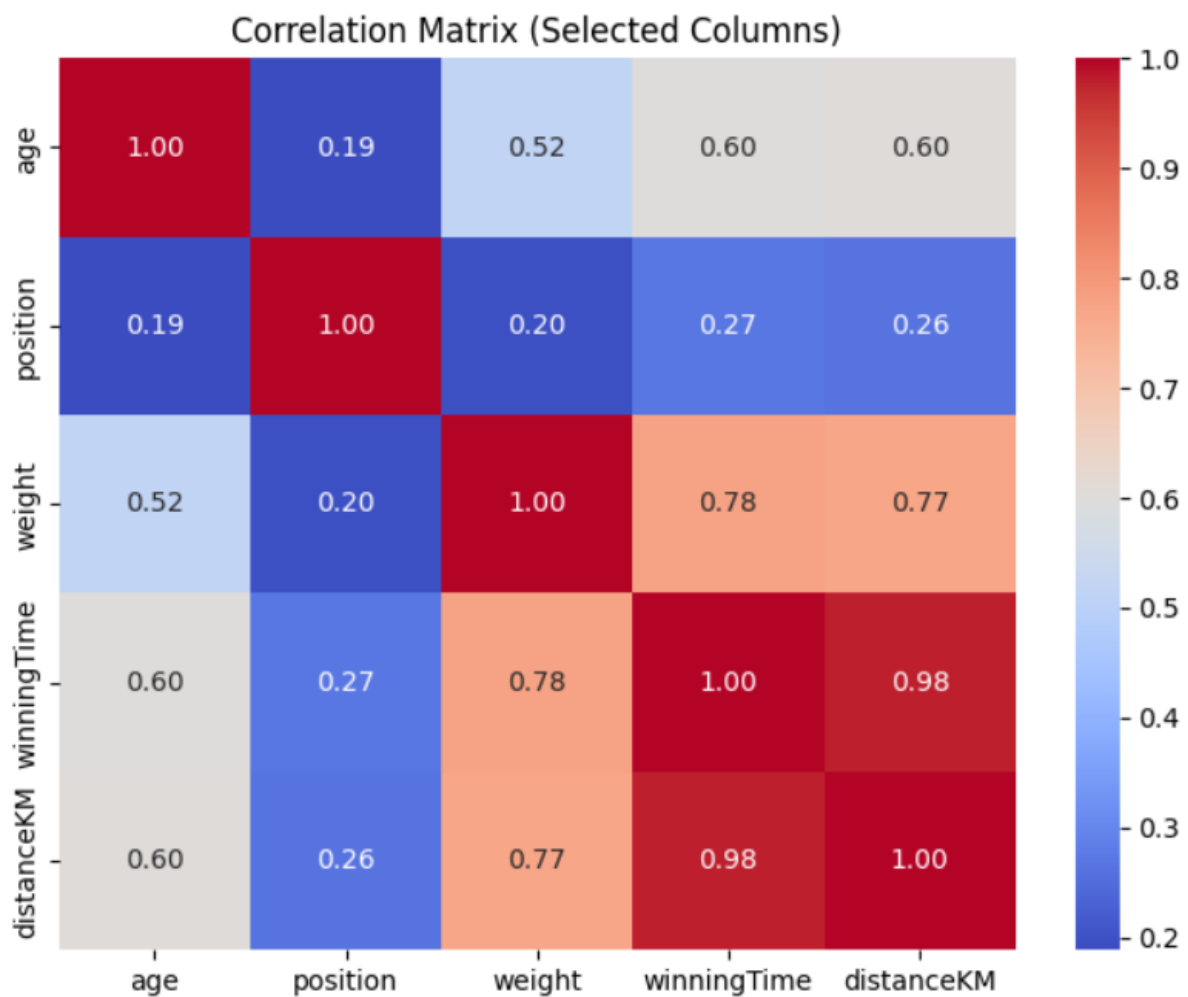
**Age, Position, WinningTime, Weight, DistanceKM**

התוצאות מצביעות על כמה קשרים חזקים בין המשתנים:

קיים קשר חזק מאוד חיובי בין **WinningTime** ל- **DistanceKM** (0.98). כלומר, ככל שהמרחק גדול יותר, זמן הזכייה נוטה להיות ארוך יותר, מה שיכול להעיד על תחרות קשה יותר במרוצים ארוכים.

קיימת קורלציה חיובית בין **Weight** ל- **DistanceKM** (0.77) ול- **WinningTime** (0.78). כלומר, משקל גבוה יותר קשור גם הוא למרחק גדול יותר ולזמן זכייה ארוך יותר.

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



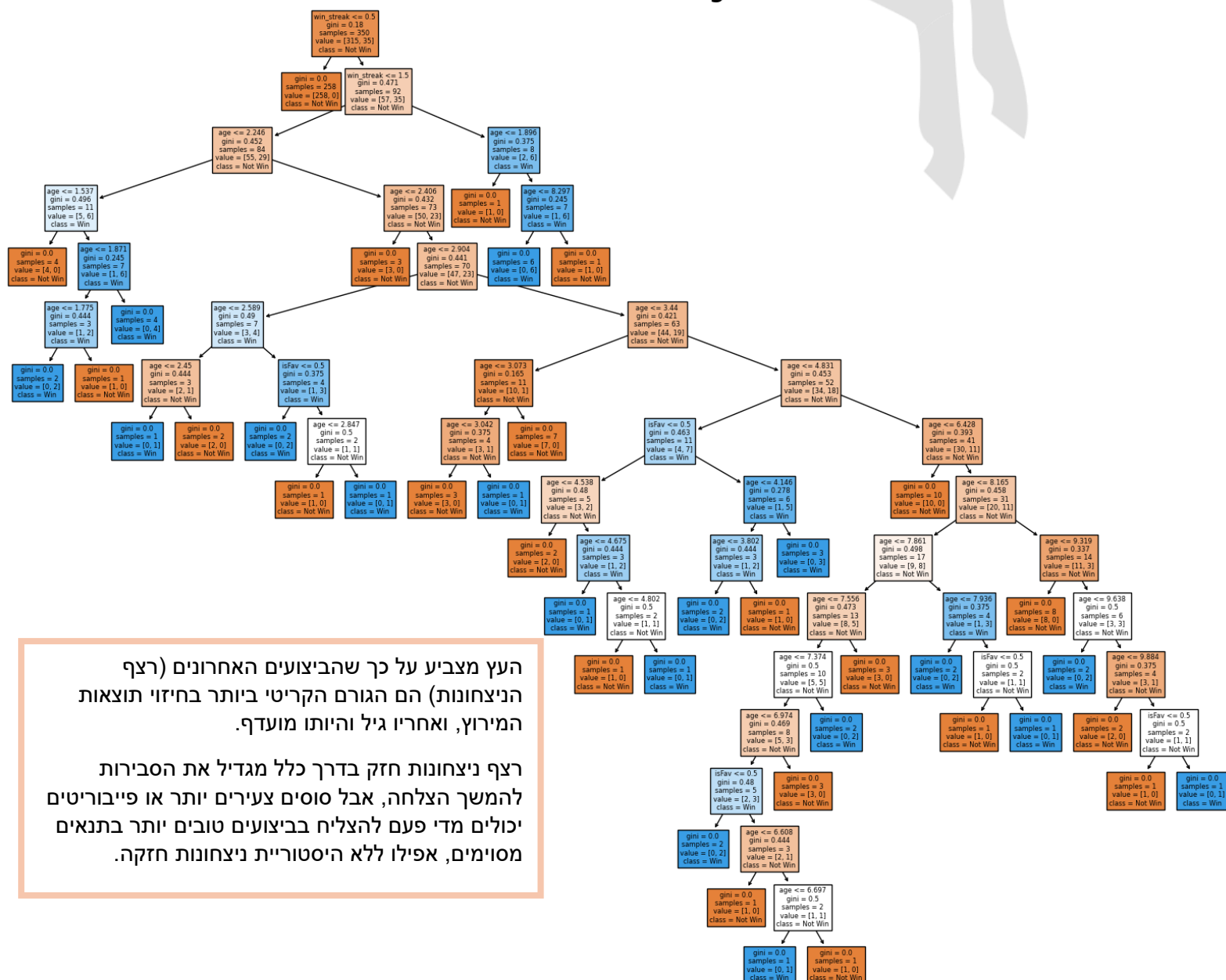
- המטריצה מראה קשר חזק בין זמן הזכייה למרחק המירוץ, וקורלציה חיובית בין משקל הסוס למרחק זמן הזכייה, כאשר סוסים מבוגרים נוטים לשקול יותר.

## ניתוח עץ החלטות:

לאחר ביצוע הניתוח בעזרת עץ החלטה לניבוי ניצחונות במרוצים, התקבלו המדדים הבאים עבור סט הבדיקה:

- **דיוק: (Accuracy)** המדד מדד את היחס בין מספר התחזיות הנכונות לסך כל התחזיות, והערך שהתקבל הוא 0.73.
- **דיוק התחזיות: (Precision)** המדד מייצג את אחוז התחזיות המדויקות של הניצחונות מתוך כלל התחזיות לניצחון, והערך שהתקבל הוא 0.72.
- **רגישות: (Recall)** המדד מייצג את אחוז הזיהוי הנכון של הניצחונות בפועל מתוך כלל הניצחונות, והערך שהתקבל הוא 0.71.
- **מטריצת בלבול: (Confusion Matrix)** מטריצת הבלבול מדגימה את התפלגות התחזיות הנכונות והטעויות, תוך חלוקה לקטגוריות של ניצחון ופספוס בניצחון.

Decision Tree for Predicting Race Wins

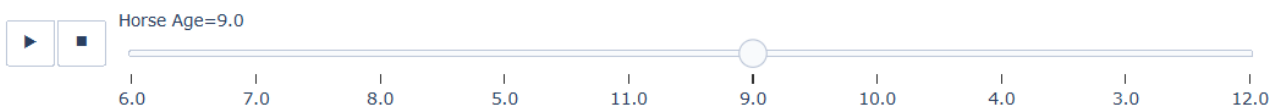
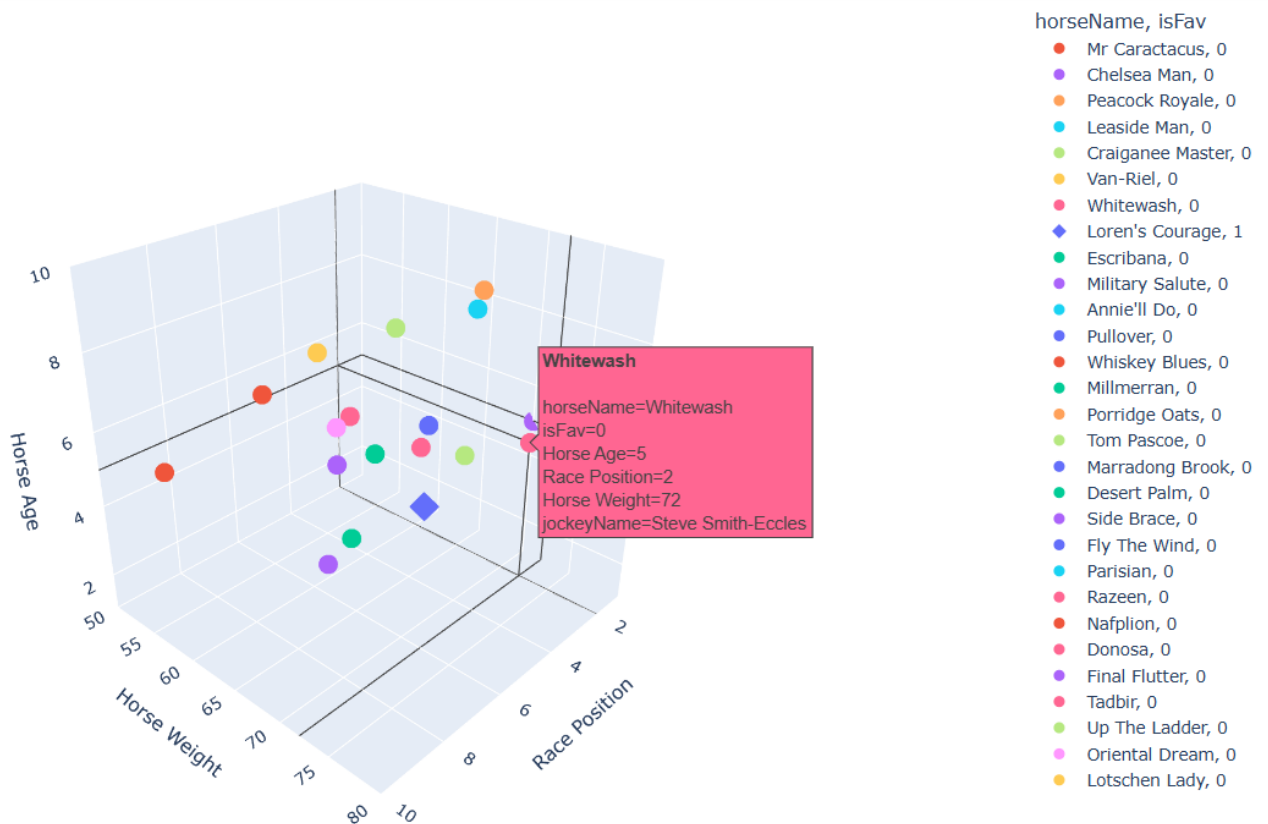


## ניתוח גרף פיזור תלת-ממדי:

הקוד יוצר עלילת פיזור תלת ממדית באמצעות **Plotly Express**, הממחיש את הקשר בין גיל הסוס, משקלו ומיקום המירוץ. כל נקודה מייצגת סוס, עם מידע נוסף זמין ברחף.

כל פריים מייצג גיל סוס אחר. זה מאפשר להתבונן כיצד הקשר בין משקל ומיקום הגזע משתנה על פני קבוצות גיל שונות.

כל סוס מיוצג על ידי צבע ייחודי, בנוסף יש סמל שמבדיל בין סוסים אהובים (**isFav**) לבין סוסים שאינם אהובים ומופיעים פרטים נוספים כגון שם הרוכב.



- הגרף התלת-ממדי המראה את הקשר בין גיל, משקל, ומיקום 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 שמצביע על כך שבשלב זה של חייהם, סוסים נמצאים בשיא כושרם להתחרות. סוסים מבוגרים יותר, במיוחד מעל גיל 10, פחות מיוצגים, מה שעשוי להעיד על ירידה ביכולות הפיזיות או העדפה של בעלי הסוסים להפסיק את השתתפותם במרוצים בשלב זה.

## **מסקנות לגבי ביצועי עשרת המאמנים המובילים:**

מהניתוח ניתן להסיק כי ישנם מאמנים שמצליחים להשיג אחוז ניצחונות גבוה למרות שמספר הסוסים שהם מאמנים קטן יחסית. הדבר יכול להעיד על אסטרטגיות אימון יעילות יותר או על בחירה ממוקדת של תחרויות בהן יש סיכוי גבוה יותר לניצחון. לעומת זאת, מאמנים אחרים משיגים אחוז ניצחונות נמוך יותר למרות שהם מאמנים מספר גדול יותר של סוסים, מה שיכול להצביע על יכולות אימונם או על השתתפות בתחרויות חסרות סיכוי.

## **מסקנות לגבי מגמות ביצוע של חמשת המאמנים המובילים לאורך השנים:**

התנודות במספר הניצחונות והמרוצים משקפות את האתגרים וההזדמנויות שמאמנים שונים מתמודדים איתם לאורך השנים. מאמנים שהצליחו לשמור על יציבות מסוימת הצליחו גם לשמר את מעמדם בצמרת לאורך זמן. ההצלחה של מאמן לאורך שנים תלויה בניהול נכון של משאבים והכנה מדוקדקת לקראת המרוצים. המאמנים שהצליחו לשמור על יציבות יחסית לאורך השנים הם אלו שכנראה נקטו בגישה שיטתית ומושכלת.

## **מסקנות לגבי אחוזי הניצחונות של מאמנים מובילים בתנאי מירוץ שונים:**

ניתן להסיק שמאמנים מסוימים מתמחים בתנאי מירוץ ספציפיים, מה שיכול להיות משמעותי בעת בחירת מאמן לסוס בתנאי מירוץ מסוימים. בתנאי מירוץ קשים, רוב המאמנים אינם מצליחים לנצח, מה שיכול להצביע על כך שתנאים אלו קשים במיוחד עבור סוסים ומאמנים כאחד.

### מסקנות לגבי ביצועי סוסים פייבוריטים מול סוסים לא פייבוריטים:

הנתונים יכולים לשמש לשיפור החלטות אסטרטגיות במרוצים, הן עבור מהמרים והן עבור מאמנים, בהתחשב בכך שסוסים פייבוריטים אכן מצליחים יותר, אך לא מבטיחים ניצחון מוחלט.

### מסקנות מהשפעת הגיל על הסתברות לניצחון במרוצים:

מהגרף ניתן להסיק שגיל הסוס הוא גורם משמעותי בקביעת ההסתברות לניצחון במירוץ. סוסים צעירים יותר, בעיקר בגילאים 2-5, מראים הסתברות גבוהה יותר לניצחון, בעוד שסוסים מבוגרים יותר מראים ירידה בהסתברות זו. המסקנה היא שיש לשים דגש על גיל הסוס כאשר מנתחים את סיכוייו לנצח במירוץ. ייתכן שזה יכול להיות כלי עזר בקבלת החלטות אסטרטגיות בעת בחירת סוסים פייבוריטים במרוצים.

### מסקנות לגבי השפעת מצב המסלול על זמן הזכייה:

ניתן להסיק שמצבי מסלול יותר 'כבדים' או 'רטובים' כמו **Soft To Heavy** ו-**Heavy** דורשים מהסוסים יותר זמן לסיים את המירוץ. בעוד שמצבי מסלול 'קלים' יותר כמו **Fast** מאפשרים זמנים מהירים יותר. תוצאה זו עשויה לשקף את הקושי הפיזי הנדרש מהסוסים כאשר הם מתמודדים עם תנאי מסלול פחות אידיאליים. ההבדלים בזמני הזכייה בין מצבי המסלול השונים עשויים להיות קריטיים עבור מאמנים, בעלי סוסים, ומהמרים, שכן הם מצביעים על כך שהתנאים המסלוליים יכולים להשפיע משמעותית על ביצועי הסוסים.

### מסקנות לגבי השפעת התפלגות תנאי המסלול לאורך השנים:

מהניתוח ניתן להסיק כי תנאי המסלול המשפיעים ביותר על תוצאות המרוצים היו אלה שהופיעו בתדירות גבוהה יותר כמו **Good** ו-**Soft**. המשמעות היא שהתנאים האלה כנראה היו הסטנדרטיים והעדיפים עבור מרבית המרוצים, מה שמעיד על יציבות מסוימת בתנאי המסלול לאורך השנים. שימוש בסולם לוגריתמי להדגשת תנאים שהיו פחות נפוצים אך עדיין חשובים, כמו **Fast, Frozen** ו-**Abandoned**, מאפשר לנו להבין את הפיזור והחשיבות של תנאים נדירים במרוצי סוסים.

### מסקנות לגבי קשרי הקורלציה בין המשתנים:

הניתוח מצביע על כך שמרחק המירוץ חזמן הזכייה הם משתנים הקשורים באופן ישיר וחזק זה לזה, מה שמדגיש את השפעת המרחק על תוצאות המירוץ. בנוסף, הקורלציה החיובית בין משקל הסוס לבין המרחק חזמן הזכייה מרמזת שמשקל רב עשוי להוות גורם מכריע במרוצים ארוכים. הקשר בין הגיל והמשקל מצביע על כך שסוסים מבוגרים נוטים להיות כבדים יותר, מה שעשוי להשפיע על הביצועים שלהם בתנאים מסוימים.

### **מסקנות ניתוח עץ החלטות:**

העץ מורכב למדי עם פיצולים רבים, מה שמצביע על כך שחזוי תוצאות להצלחת הסוס מושפע ממספר גורמים. צמתים מסוימים בעץ מובילים לתחזיות בעלות ביטחון גבוה (צמתים כתומים או כחולים), בעוד שלאחרים יש תוצאות מעורבות יותר. ניתן להשתמש בעץ ההחלטות הזה כדי לנתח ולחזות את תוצאות המירוץ על סמך נתוני עבר של הסוס, ולהיעזר ולזהות גורמי מפתח שתורמים לניצחון.

### **מסקנות עבור גרף פיזור תלת-ממדי:**

ההדמיה מספקת מבט מקיף על משחק הגומלין המורכב בין גיל הסוס, משקלו וביצועי המירוץ. זה מדגיש את האופי הרב-גוני של מרוצי סוסים, שבו גורמים רבים תורמים להצלחת הסוס מעבר לתכונות פיזיות כמו גיל ומשקל.

