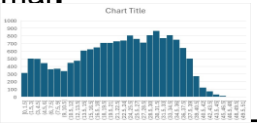
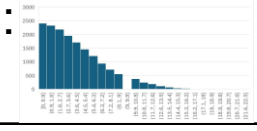
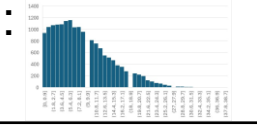
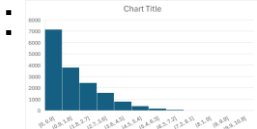


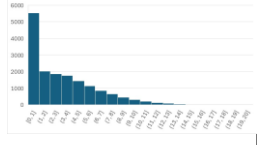
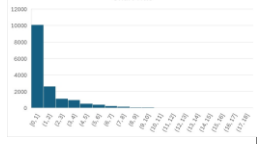

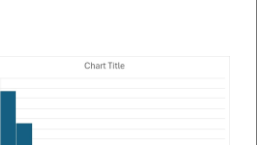
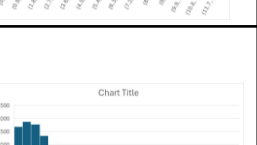
NBA - Player Stats - Season 24/25

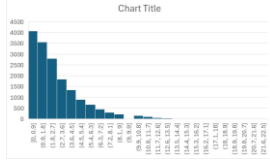
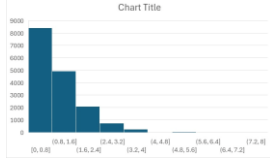
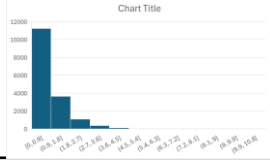
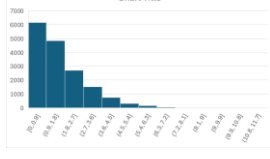
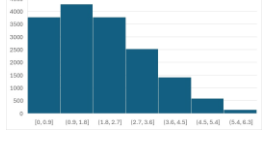
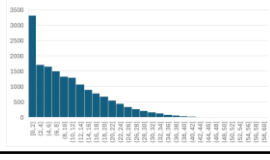
Data Analysis Project

Tamir Ovadia

טבלה מסכמת של כלל המשתנים

Variable Name	Representation	Type	Range	Missing Values	First Distribution	Dependent \ Independent
Tm	Abbreviation of the player's team	categorical	-	-	no meaning (cat)	independent
Res	Result of the game for the player's team	Binary	{0,1}	-	no meaning (bin)	independent
MP	Minutes played	float	[0-50.48]	-	Normal: 	independent
FG	Field goals made.	int	[0-22]	-	Ln: 	independent
FGA	Field goal attempts	int	[0-38]	-	Ln: 	independent
3P	3-point field goals made.	int	[0-10]	-	Ln: 	independent

Variable Name	Representation	Type	Range	Missing Values	First Distribution	Dependent \ Independent
3PA	3-point field goal attempts.	int	[0-20]	-	Ln 	independent
FT	Free throws made.	int	[0-18]	-	Ln 	independent
FTA	Free throw attempts.	int	[0-26]	-	Ln 	independent
ORB	Offensive rebounds	int	[0-12]	-	Ln 	independent
DRB	Defensive rebounds	int	[0-23]	-	Ln 	independent

Variable Name	Representation	Type	Range	Missing Values	First Distribution	Dependent \ Independent
AST	Assists	[0-22]	int	-		independent
STL	Steals	[0-8]	int	-		independent
BLK	Blocks	[0-10]	Int	-		independent
TOV	Turnovers	[0-11]	int	-		independent
PF	Personal fouls.	[0-6]	int	-		independent
PTS	Total points scored.	[0-60]	int	-		dependent

מפה תרמית + הקשר בין משתנה x_j לבין משתנה x_j

משתנה המטרה שלנו הוא $PTS - Y$ שמייצג את מספר הנקודות שיקלע שחקן.

PTS	PF	TOV	BLK	STL	AST	DRB	ORB	FTA	FT	3PA	3P	FGA	FG	MP	
														1.000	MP
													1.000	0.727	FG
												1.000	0.894	0.793	FGA
											1.000	0.584	0.610	0.490	3P
										1.000	0.817	0.738	0.580	0.602	3PA
									1.000	0.309	0.239	0.534	0.499	0.481	FT
								1.000	0.958	0.276	0.214	0.535	0.510	0.488	FTA
							1.000	0.190	0.148	-0.072	-0.068	0.225	0.249	0.288	ORB
						1.000	0.380	0.357	0.324	0.188	0.149	0.462	0.473	0.563	DRB
					1.000	0.336	0.077	0.369	0.372	0.391	0.303	0.545	0.483	0.565	AST
				1.000	0.268	0.185	0.089	0.190	0.183	0.226	0.176	0.314	0.285	0.372	STL
			1.000	0.083	0.063	0.314	0.234	0.148	0.130	0.021	0.011	0.164	0.186	0.248	BLK
		1.000	0.111	0.210	0.432	0.332	0.117	0.356	0.349	0.310	0.244	0.469	0.428	0.471	TOV
	1.000	0.265	0.153	0.164	0.208	0.275	0.209	0.200	0.187	0.185	0.156	0.296	0.282	0.424	PF
1.000	0.282	0.446	0.172	0.286	0.503	0.456	0.209	0.655	0.662	0.647	0.683	0.895	0.969	0.742	PTS

סולם צבעים

קשר חזק
בין 0.6 ל-1 או בין -0.6 ל-1
קשר בינוני
בין 0.3 ל-0.6 או בין -0.3 ל-0.6
קשר חלש
בין -0.3 ל-0.3

מולטיקולינאריות והשערות ראשוניות

במצב של מולטיקולינאריות, שני משתנים בלתי תלויים או יותר במודל רגרסיה לינארית, מקיימים ביניהם קשר ליניארי חזק, ולכן מספקים מידע חופף. הדבר מקשה על זיהוי ההשפעה הייחודית של כל משתנה על משתנה המטרה ופוגע בדיוק ההערכות. בקובץ הנתונים קיים חשש למולטיקולינאריות עקב קורלציה גבוהה בין משתנים בלתי תלויים. בין המשתנים המתארים הצלחה וניסיונות (Attempts) קיימת קורלציה גבוהה. לדוגמא מתאם של 0.89 בין המשתנים FG ו-FGA, מכיוון שמספר הסלים (FG) תלוי במספר הניסיונות (FGA). לכן, מומלץ להסיר אחד מהמשתנים בהתאם לקשר שלו עם משתנה המטרה (PTS).

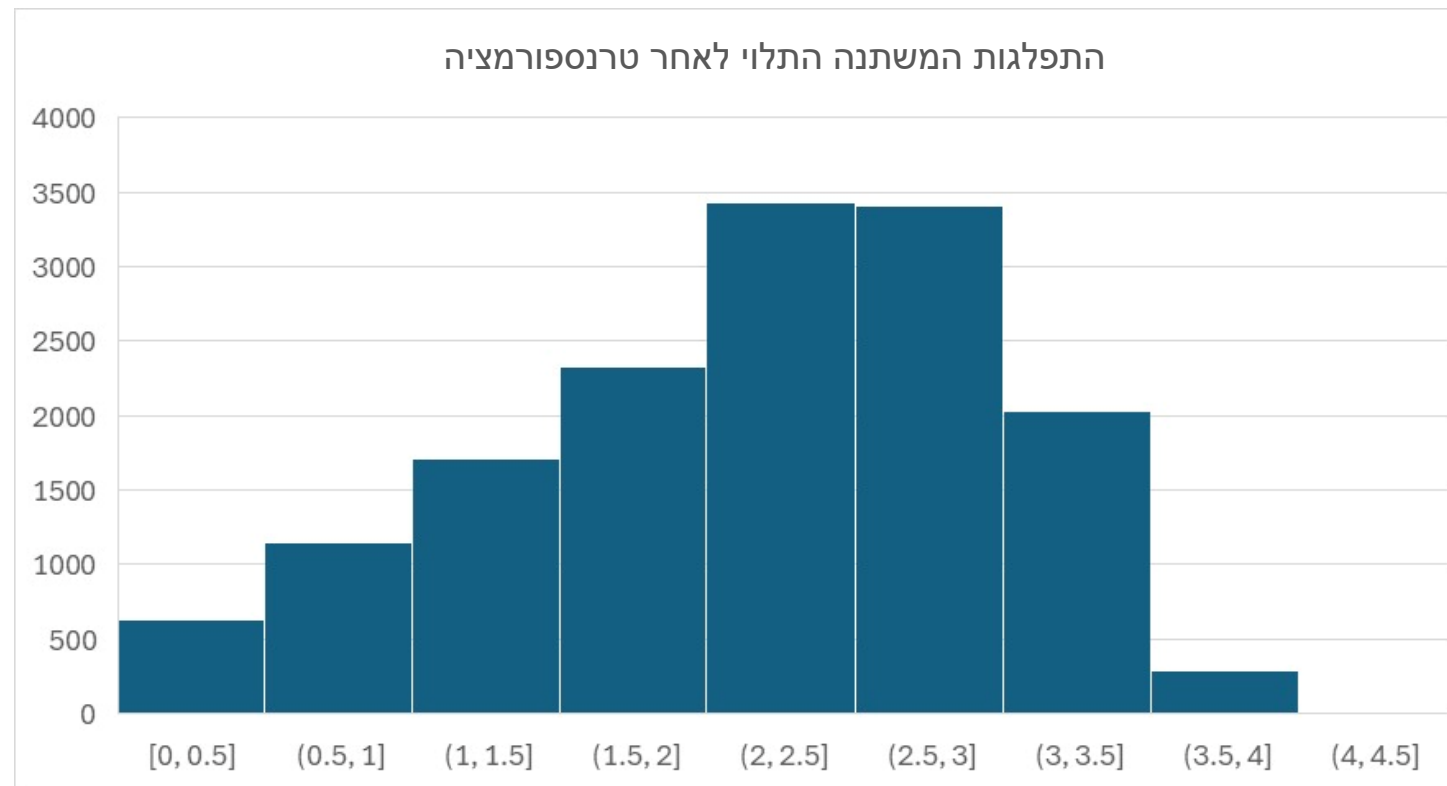
משתנים בלתי תלויים עם מתאם גבוה (מעל 0.6) עם משתנה המטרה (התלוי) יכולים לנבא אותו. בקובץ הנתונים, FG (0.968), 3P (0.68) ו-MP (0.74) הם דוגמאות למשתנים בעלי קשר חזק עם משתנה המטרה.

השערות ותובנות ראשונות מתוך חיתוך הנתונים:

1. ככל ששחקן משחק יותר דקות (MP) כך גם מספר הנקודות שלו (PTS) יהיה גבוה יותר ולהיפך (מתאם 0.74)
2. אין קשר בין יכולות הגנתיות (BLK ו-STL) למספר הנקודות (PTS) (מתאם 0.17 ו-0.28)
3. יש קשר חזק בין כל המשתנים המתארים ניסיונות (ATTEMPTS) לבין כל המשתנים המתארים הצלחות. לדוגמא: FG ו-FGA (מתאם 0.89), 3P ו-3PA (מתאם 0.81) וכו'. לכן, נשקול להוריד חלק מהמשתנים הללו.

המרה להתפלגות נורמלית:

משתנה המטרה (PTS) מתפלג LN. נערכה טרנספורמציה כדי להתאים אותו להתפלגות נורמלית, כפי שנלמד.



צמצום משתנה קטגוריאלי:

כדי לצמצם את 30 הקטגוריות במשתנה הקטגוריאלי TM (קבוצת שחקן), השתמשנו בפונקציה מקשרת. לאחר ניסוי עם מספר פונקציות מקשרות ופקטורים, הפקטור 0.1 נבחר מכיוון ששמר על הרעש בנתונים באופן סביר. התהליך שבוצע:

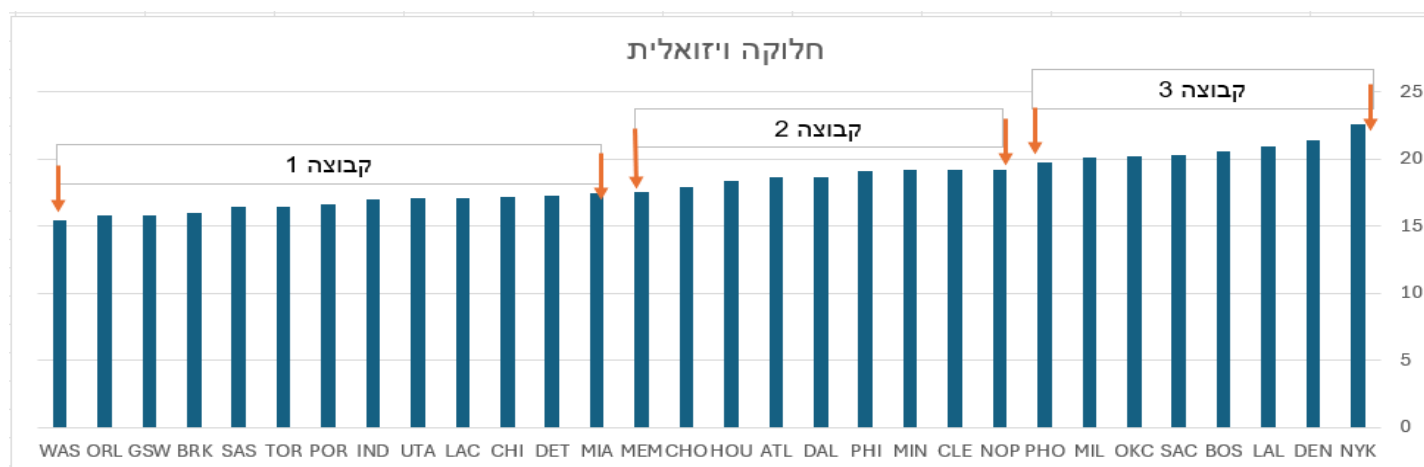
1. יצירת Pivot Table עם המשתנה TM וממוצע המשתנה התלוי PTS.

2. חישוב השונות (varp) של PTS עבור כל קבוצה.

3. חישוב הפונקציה המקשרת: $\mu + 0.1 * varp$

4. מיון התוצאות.

5. הצגת גרף התוצאות וחלוקה ויזואלית ל-3 קבוצות.



חלוקה לקבוצות

values	פונקציה מקשרת	TM
3	22.60793255	NYK
3	21.36855676	DEN
3	20.92731	LAL
3	20.57702596	BOS
3	20.27502808	SAC
3	20.19939566	OKC
3	20.15347347	MIL
3	19.7839571	PHO
2	19.25168379	NOP
2	19.24557853	CLE
2	19.21707469	MIN
2	19.06997145	PHI
2	18.67032843	DAL
2	18.65661989	ATL
2	18.3540046	HOU
2	17.90476045	CHO
2	17.59982639	MEM
1	17.44233713	MIA
1	17.26394534	DET
1	17.15402177	CHI
1	17.06279464	LAC
1	17.06001369	UTA
1	17.01519758	IND
1	16.60385571	POR
1	16.45233119	TOR
1	16.44357111	SAS
1	15.98612115	BRK
1	15.80984323	GSW
1	15.78232889	ORL
1	15.47791693	WAS

לאחר החלוקה הויזואלית לקבוצות, הוקצו להן הערכים 1, 2 ו-3.

באמצעות VLOOKUP הערכים הועברו לעמודה חדשה בשם Tm value בדאטה המקורי, ואומתה התאמה מדויקת בין הקבוצה לערך.

דגמנו מספר רשומות בדאטה המקורי ובדקנו שאכן הקבוצה מתאימה לערך שהגדרנו עבורה ושההמרה נעשתה בצורה מדויקת.

ליצירת one-hot vector, יצרנו שתי עמודות חדשות: Tm group 2 ו-Tm group 3. העמודה Tm value סוננה לפי ערך 2, וברשומות המתאימות הוצב הערך 1 בעמודה Tm group 2. פעולה דומה בוצעה עבור ערך 3 והעמודה Tm group 3. בזאת סיימנו את העבודה עם המשתנה הקטגוריאלי- יש לנו 2 עמודות שמייצגות את 3 הערכים של הקבוצות.

Tm group 3	Tm group 2
1	0
0	1
0	0

קבוצות עם הערך 3

קבוצות עם הערך 2

קבוצות עם הערך 1

ביצוע מבחן T:

ביצענו מבחן T עבור המשתנה Tm- המייצג את הקבוצה של שחקן. הקטנו את מספר הקבוצות מ-30 ל-3 סוגים (כמפורט קודם). הדאטה סונן שלוש פעמים, פעם אחת לכל קבוצה (1, 2, 3) ולקחנו את המשתנה התלוי של כל ערך. משום שבביצוע מבחן T יש לבדוק שהנתונים מתפלגים נורמלית, לקחנו את המשתנה התלוי לאחר ההמרה שביצענו למשתנה התלוי.

כדי למנוע הטיה בתוצאות מבחן T עקב כמות גדולה של נתונים, נדגמו באקראי 1,000 רשומות באמצעות הפונקציה RANDBETWEEN. לאחר מכן, ביצענו מבחן T ברמת מובהקות של 0.05 על 1,000 הרשומות שנדגמו, כדי לבדוק את ההבדלים המובהקים בין הקבוצות.

t-Test: Two-Sample Assuming Equal Variances		
LN pts tm group2	LN pts tm group1	
2.002575488	1.912188184	Mean
1.020286343	0.975113519	Variance
1000	1000	Observations
	0.997699931	Pooled Variance
	0	Hypothesized Mean Difference
	1998	df
	-2.023449947	t Stat
	0.021579836	P(T<=t) one-tail
	1.64561663	t Critical one-tail
	0.043159673	P(T<=t) two-tail
	1.961152015	t Critical two-tail

לפי ה- PVALUE, ניתן לומר כי יש הבדל בין שתי הקבוצות 1 ו-2, בר"מ 0.05.

המסקנה היא- יש השפעה לסוג הקבוצה (קבוצה מסוג 1 או 2) למספר הנקודות ששחקן ספציפי קלע.

t-Test: Two-Sample Assuming Equal Variances		
LN pts tm group3	LN pts tm group2	
1.96960062	2.002575488	Mean
1.234919756	1.020286343	Variance
1000	1000	Observations
	1.12760305	Pooled Variance
	0	Hypothesized Mean Difference
	1998	df
	0.694368402	t Stat
	0.243765967	P(T<=t) one-tail
	1.64561663	t Critical one-tail
	0.487531933	P(T<=t) two-tail
	1.961152015	t Critical two-tail

לפי ה-PVALUE, ניתן לומר כי אין הבדל בין שתי הקבוצות 2 ו-3, בר"מ 0.05.

המסקנה היא- אין השפעה לסוג הקבוצה (קבוצה מסוג 2 או 3) למספר הנקודות ששחקן ספציפי קלע.

t-Test: Two-Sample Assuming Equal Variances		
LN pts tm group3	LN pts tm group1	
1.96960062	1.912188184	Mean
1.234919756	0.975113519	Variance
1000	1000	Observations
	1.105016637	Pooled Variance
	0	Hypothesized Mean Difference
	1998	df
	-1.221255605	t Stat
	0.111066671	P(T<=t) one-tail
	1.64561663	t Critical one-tail
	0.222133342	P(T<=t) two-tail
	1.961152015	t Critical two-tail

לפי ה-PVALUE, ניתן לומר כי אין הבדל בין שתי הקבוצות 1 ו-3, בר"מ 0.05.

המסקנה היא- אין השפעה לסוג הקבוצה (קבוצה מסוג 2 או 3) למספר הנקודות ששחקן ספציפי קלע.

הרצת הרגרסיה:

לצורך הרגרסיה, בחרנו 9 משתנים בלתי תלויים ומשתנה תלוי אחד מתוך כלל הנתונים. בחירת 9 המשתנים הבלתי תלויים נעשתה בהתאם להנחיות (לפחות משתנה קטגוריאלי אחד, משתנה בינארי אחד ומשתנה רציף אחד) ובהתבסס על ניתוח מקדים שכלל בדיקת קורלציה ומולטיקולינאריות. המשתנים שנבחרו הראו קורלציה גבוהה עם המשתנה התלוי, אך לא הציגו מולטיקולינאריות גבוהה בינם לבין משתנה בלתי תלוי אחר.

כפי שנותח קודם, משתנים המייצגים ניסיונות ATTEMPTS הראו קורלציה גבוהה עם משתנים המייצגים הצלחות (FG ו-FGA, 3P ו-3PA, FT ו-FTA), לכן, לצורך הרגרסיה, נבחרו רק משתני ההצלחות (**FT**, **3P**, **FG**).

שאר המשתנים שנכללו ברגרסיה הם:
משתנה קטגוריאלי - **Tm** המייצג את קבוצת השחקן.
משתנה בינארי - **Res** המייצג את תוצאת המשחק (1 לניצחון הקבוצה, 0 להפסד).
משתנים רציפים - **TOV**, **AST**, **DREB**, **MP** שהראו את הקורלציה הגבוהה ביותר עם המשתנה התלוי מבין המשתנים הנותרים.

תוצאות:

טיב המודל: המודל הוא מודל טוב עם R Square גבוה- 0.85. ניתן לראות שאין OVERFITTING ($R^2 > 0.95$)

Regression Statistics	
0.922190584	Multiple R
0.850435473	R Square
0.850344833	Adjusted R Square
0.404138924	Standard Error
16512	Observations

כל המשתנים ברגרסיה נמצאו מובהקים, ככל הנראה כתוצאה מבחירה מושכלת של המשתנים וניתוח מקדים איכותי, אך יש לזכור שהדבר יכול לנבוע גם מכמות הנתונים הגדולה, אשר עשויה להטות את ערך ה-PVALUE.

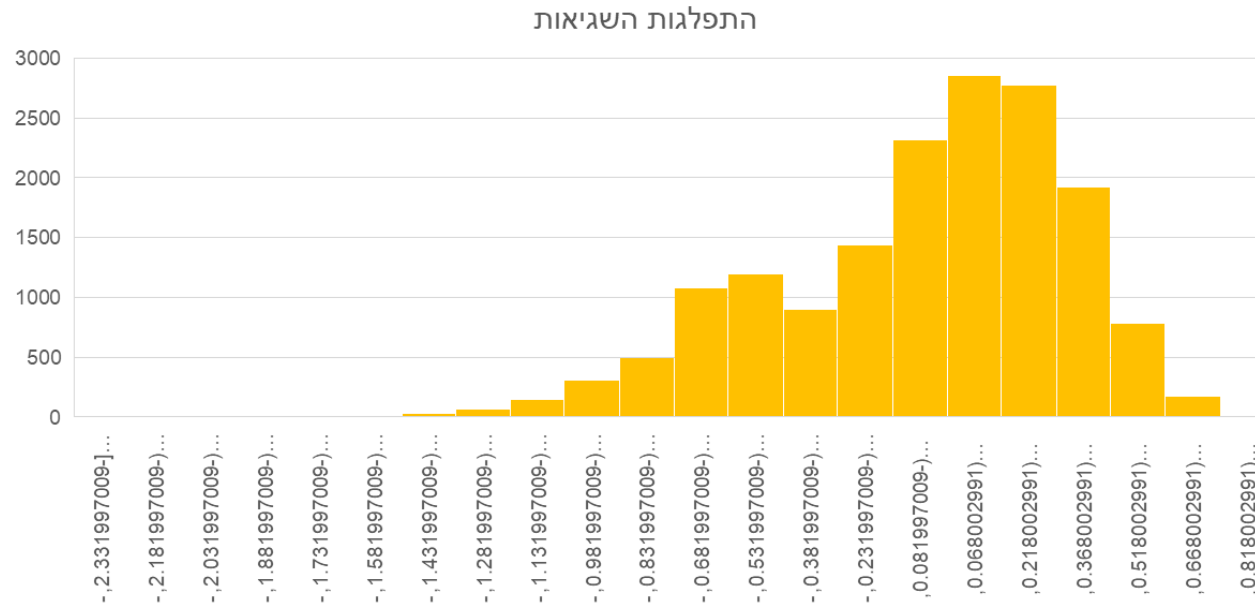
Upper 95.0%	Lower 95.0%	Upper 95%	Lower 95%	P-value	t Stat	Standard Error	Coefficients
0.508	0.474	0.508	0.474	0.000	56.818	0.009	0.491 Intercept
-0.025	-0.055	-0.025	-0.055	0.000	-5.349	0.008	-0.040 Tm group 2
-0.110	-0.141	-0.110	-0.141	0.000	-16.002	0.008	-0.126 Tm group 3
0.027	0.001	0.027	0.001	0.029	2.187	0.006	0.014 RES
0.030	0.028	0.030	0.028	0.000	58.336	0.000	0.029 MP
0.175	0.168	0.175	0.168	0.000	103.144	0.002	0.171 FG
0.099	0.089	0.099	0.089	0.000	34.547	0.003	0.094 3P
0.065	0.058	0.065	0.058	0.000	36.337	0.002	0.061 FT
-0.011	-0.017	-0.011	-0.017	0.000	-9.514	0.002	-0.014 AST
0.006	0.000	0.006	0.000	0.029	2.179	0.001	0.003 DRB
-0.007	-0.018	-0.007	-0.018	0.000	-4.775	0.003	-0.012 TOV

משוואת הרגרסיה:

$$Y = 0.491 - 0.04Tm2 - 0.126Tm3 + 0.014RES + 0.029MP + 0.171FG + 0.094(3P) + 0.061FT - 0.014AST + 0.003DRB - 0.012TOV$$

סעיף c:

ניתוח השגיאות מראה שהתפלגותה- Residuals נורמלית. לכן, בהתבסס על ניתוח התוצאות, ניתן להסיק שהמודל שלנו לחיזוי מספר הנקודות ששחקן יקלע במשחק הוא מודל טוב.



סעיף d:

תאימות המודל טובה, כפי שמצביע ערך ה- R Square של 0.85. אין חשש ל- Overfitting מאחר ש- R Square אינו גבוה מ-0.95.

לגבי מולטיקולינאריות, בניתוח המקדים זיהינו משתנים עם קורלציה גבוהה ונמנענו מהכללתם יחד ברגרסיה. עם זאת, טבלת הקורלציות מראה כי המשתנה MP (שנכלל ברגרסיה) בעל קורלציה בינונית עד גבוהה עם מספר משתנים אחרים ברגרסיה (FT, FG, TOV, AST). למרות שהכללנו את MP בשל קורלציה גבוהה עם משתנה המטרה (ובהתאם לדרישה ל-9 משתנים), נציע לבחון מחדש את המודל ללא משתנה זה כדי לבדוק אם הביצועים משתפרים.

רגרסיה לוגיסטית:

ביצענו את הרגרסיה על המשתנה הבינארי Res המייצג תוצאת משחק (1- הקבוצה של השחקן ניצחה, 0- הפסידה). המשתנה הרציף שבחרנו הוא MP המייצג את מספר הדקות ששחקן שיחק במשחק.

המקדמים שקיבלנו לאחר הרצת הרגרסיה הלוגיסטית הם:

B0	0.000958864
B1	-3.43257E-05

סעיף b:

נוסחת הרגרסיה היא:

$$Y = 0.00095 - 3.4 * 10^{-5}X$$

Pivot table למשתנים קטגוריאליים:

בחרנו להציג את המשתנה **Tm** ואת המשתנה הבינארי **Res** (משתנה בינארי הוא בעצם מקרה פרטי של משתנה קטגוריאלי)

עבור המשתנה Res (Win=1 ,Lose=0):

StdDevp of PTS	Average of PTS	Res
8.38	9.93	0
9.22	11.11	1

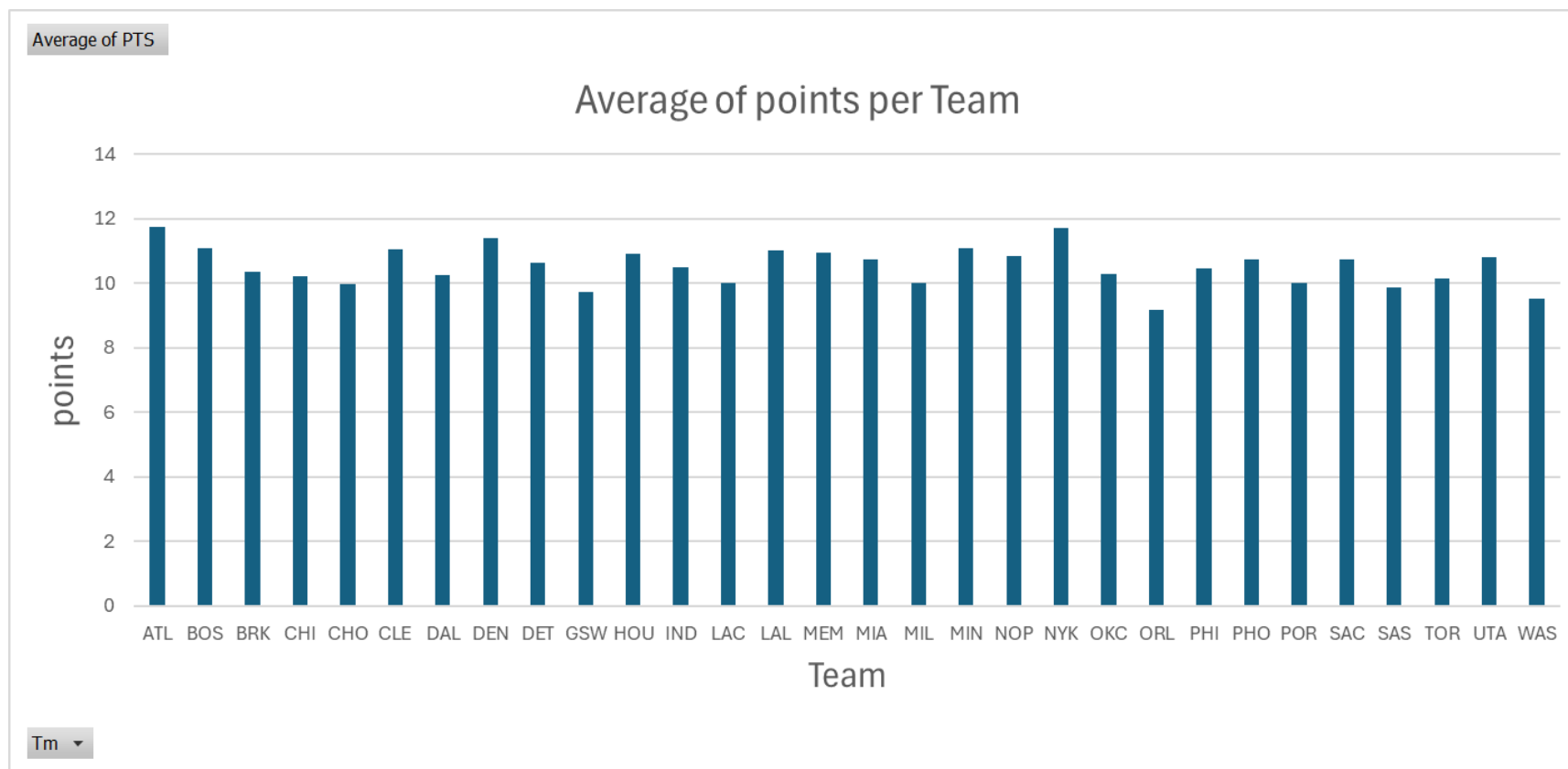
עבור המשתנה Tm:

StdDevp of PTS	Average of PTS	Tm
8.30	11.76	ATL
9.73	11.11	BOS
7.50	10.37	BRK
8.32	10.23	CHI
8.89	10.00	CHO
9.04	11.07	CLE
9.18	10.25	DAL
9.98	11.40	DEN
8.14	10.64	DET
7.80	9.72	GSW
8.61	10.93	HOU
8.07	10.50	IND
8.39	10.03	LAC
9.95	11.03	LAL
8.15	10.96	MEM
8.18	10.75	MIA
10.06	10.03	MIL
9.02	11.08	MIN
9.17	10.83	NOP
10.43	11.72	NYK
9.95	10.30	OKC
8.13	9.17	ORL
9.28	10.46	PHI
9.51	10.73	PHO
8.12	10.02	POR
9.76	10.74	SAC
8.11	9.87	SAS
7.93	10.16	TOR
7.90	10.82	UTA
7.72	9.52	WAS

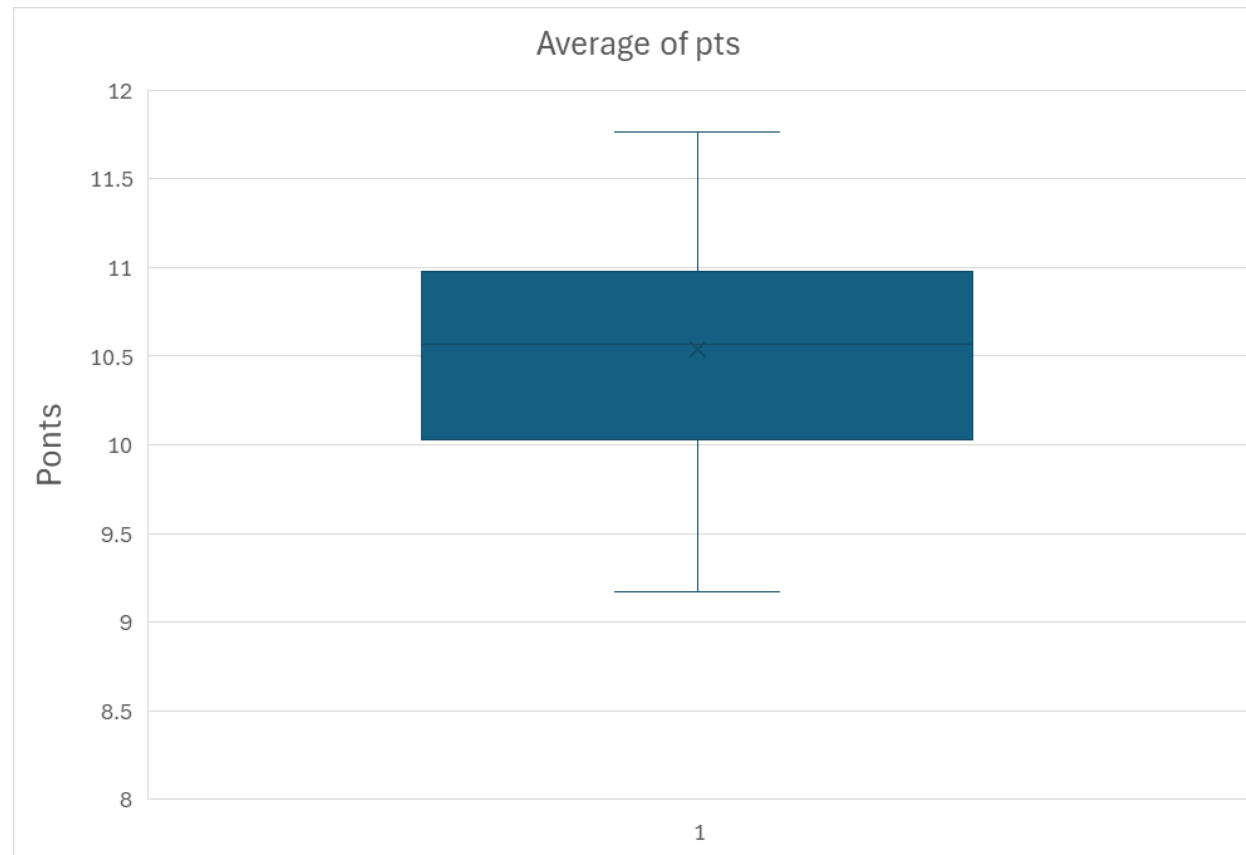
ויזואליזציה:

גרף מסוג **Bars** מתאים להשוואה בין קטגוריות או קבוצות שונות.

עושים שימוש בגרף זה כי מאפשר לראות בקלות את ההבדלים בין ממוצעי הקבוצות. אנחנו השונו באמצעות גרף זה את ממוצעי הנקודות (PTS) עבור כל קבוצה.

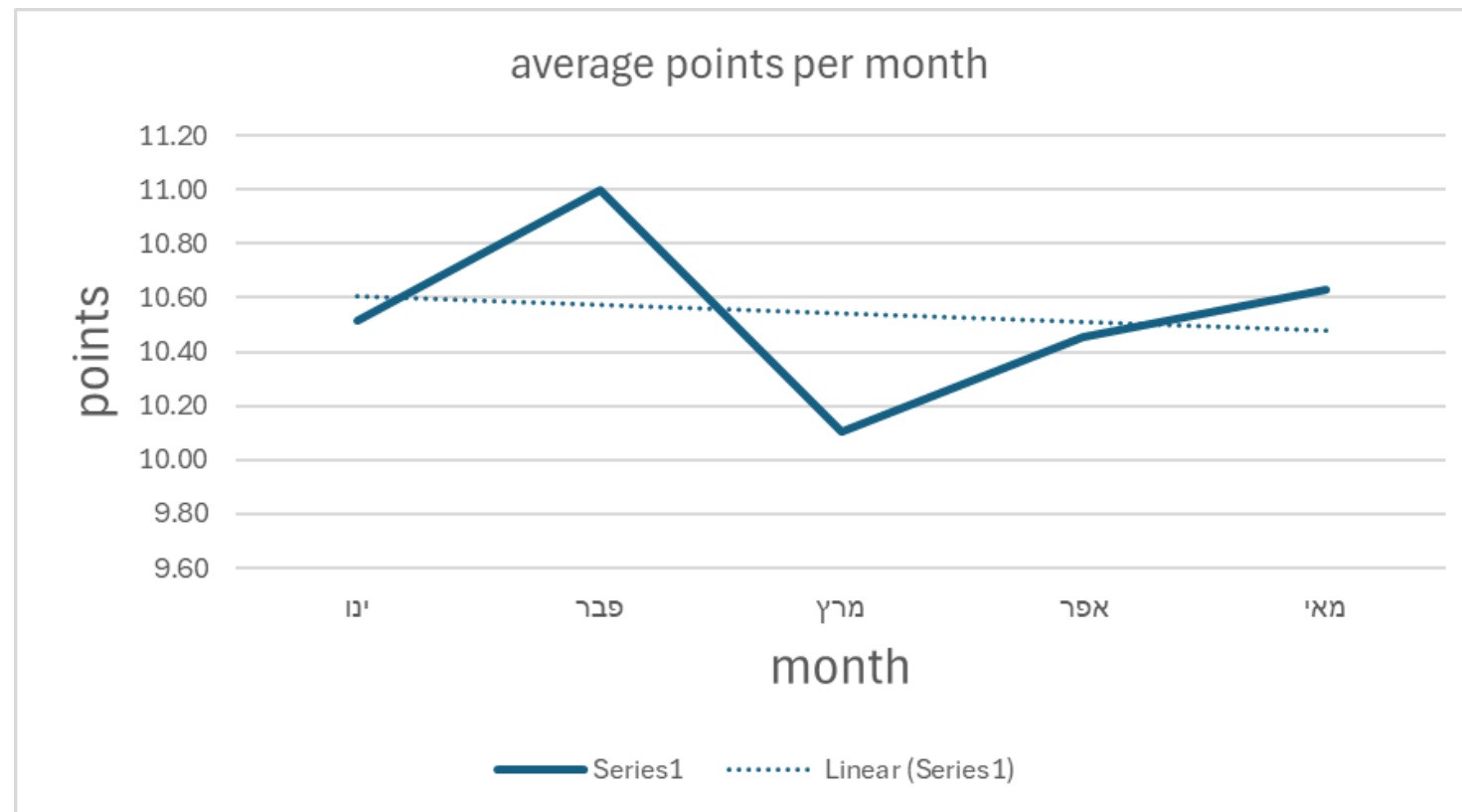


גרף מסוג **Box-Plot** מתאים להשוואת חציונים ופיזור נתונים, ולאיתור חריגים (במיוחד כאשר הנתונים מתפלגים נורמלית). אנחנו השונו בין הממוצעים של ערכי המשתנה התלוי (PTS) עבור כל קבוצה.

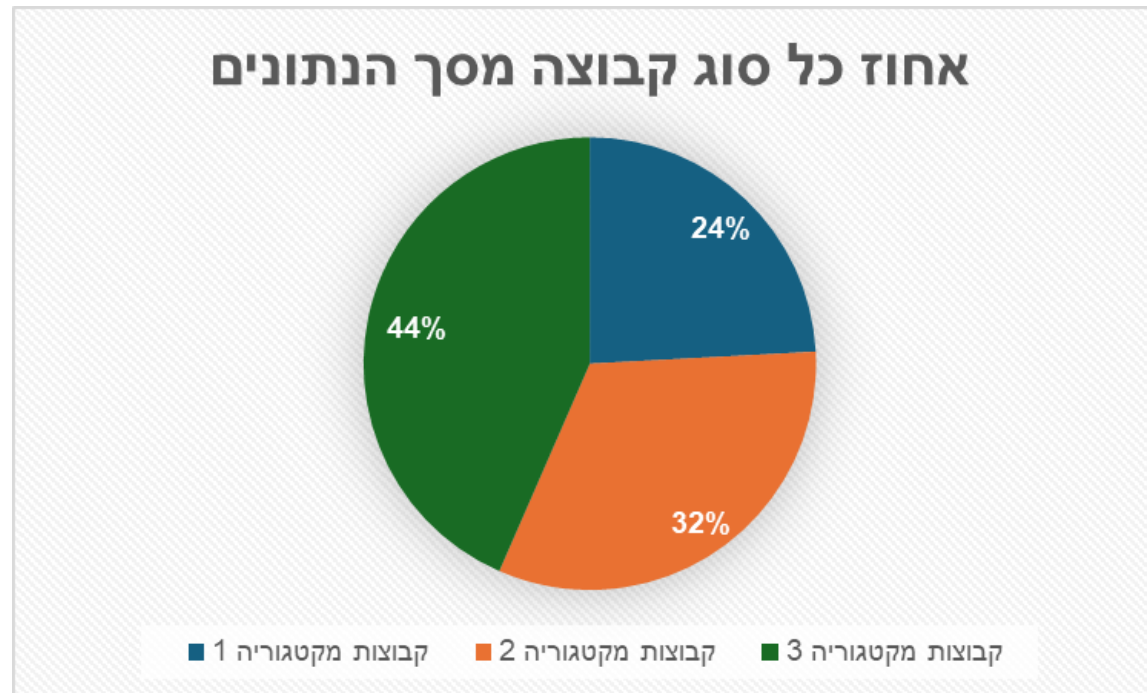


גרף מסוג **Line-Plot** משמש להצגת מגמות לאורך זמן.

השווינו את ממוצעי הנקודות שזה המשתנה התלוי (PTS) לפי חודש מתחילת השנה. ניתן לראות מגמת ירידה קלה בממוצע הנקודות ככל שהחודשים מתקדמים.



גרף מסוג **Pie Chart** משמש להצגת הפרופורציה של כל קבוצה ביחס לשלם (סך הנתונים).
בניתוח שלנו, השתמשנו בו כדי להראות את החלק היחסי של כל אחת **מסוגי הקבוצות השונות** (לאחר הצמצום שביצענו בחלק השני) מתוך סך הנתונים.



תובנות עיקריות מניתוח הנתונים:

1. **מגרף ה- Line Plot** נצפית מגמת ירידה כללית בכמות הנקודות ששחקני ה- NBA קולעים לאורך החודשים. ייתכן שהדבר נובע מעייפות מצטברת כתוצאה מעומס המשחקים במהלך העונה. בתחילת העונה השחקנים רעננים יותר, ועם התקדמות העונה הם מתעייפים.
2. **מניתוח ה- Bar-Plot** (השוואת נקודות בין קבוצות) לא נמצאו הבדלים משמעותיים בממוצע הנקודות בין קבוצות השחקנים השונות. ייתכן שהדבר נובע מכך שגם בקבוצות חלשות ישנם שחקנים בעלי יכולת קליעה גבוהה, וגם בקבוצות חזקות ישנם שחקנים שקולעים פחות. מכיוון שהדאטה כולל את כל השחקנים ששיחקו בשנה ואת סך הנקודות שלהם, הממוצעים פר קבוצה נוטים להיות דומים יחסית.
3. **מה- Line plot** החל מחודש **מרץ** נראית מגמת עלייה חוזרת בכמות הנקודות של שחקנים. הדבר מתיישב עם ההערכה שהשחקנים מגבירים את קצב המשחק והמאמץ לקראת סיום העונה הסדירה ולתחילת ה- PlayOff שמתחיל בחודש אפריל.