



פָרָוִיקָט סִינְסִיּוֹ
מַדְעֵה הַנְּתָנִים

S&P 500

דוד דסה ויצחק יוסף



S&P 500

שאלת המחקר

האם באמצעות תוצאות של עובדי
החברה מדור אחר באתר Glassdoor ניתן
לchezות את הביצועים הכלכליים של
החברות שנמצאות ב-S&P 500?



הסבר כללי על הפרויקט

אתר **Glassdoor** הוא אתר בו עובדים רבים של חברות שונות מכל העולם נרשמים וכותבים תשובות על חברותיהם. בתגובה אלה יש נתונים רבים (צינויים שהעובדים נותנים לאספקטים שונים בחברה, תשובות מילוליות על החברה וכו'). נתונים אלה יכולים לשמש כמידע פנימי ולהעיד על המצב החברתי בחברה ועל דעתם של עובדי החברה עליה. מטרת המחקר היא לבדוק האם נתונים אלה יכולים לחזות גם את הביצועים הכלכליים של החברה.



crawling

סיקת נתונים היא שיטה הכללת כריית נתונים
מקורות אינטרנט שונים. סיקת נתונים דומה מאוד
למה שעושים מנג'י החיפוש הגדולים. במלחים פשוטות,
סיקת נתונים היא שיטה למציאת קישורים באינטרנט
וקבלת מידע מהם.



מקורות מידע

GLASSDOOR

שימוש ב- Crawling
Glassdoor להשגת תגבות מ-

FINANCIAL MODELING PREP - API

שימוש ב- API להשגת נתונים
פיננסיים מ- Financial modeling .prep

WIKIPEDIA - S&P 500

שימוש בסקריפינג להעתיקת
פרטים על 500 החברות

השתמשנו בscriping כדי ללקח את טבלת S&P 500 שבה

קיימים כל הערכים על החברות כגן:

Symbol	Security	GICS Sector	GICS Sub-Industry	Headquarters Location	Date added	CIK	Founded
MMM	3M	Industrials	Industrial Conglomerates	Saint Paul, Minnesota	1957-03-04	0000066740	1902
AOS	A. O. Smith	Industrials	Building Products	Milwaukee, Wisconsin	2017-07-26	0000091142	1916
ABT	Abbott	Health Care	Health Care Equipment	North Chicago, Illinois	1957-03-04	0000001800	1888
ABBV	AbbVie	Health Care	Pharmaceuticals	North Chicago, Illinois	2012-12-31	0001551152	2013 (1888)
ACN	Accenture	Information Technology	IT Consulting & Other Services	Dublin, Ireland	2011-07-06	0001467373	1989
ATVI	Activision Blizzard	Communication Services	Interactive Home Entertainment	Santa Monica, California	2015-08-31	0000718877	2008
ADM	ADM	Consumer Staples	Agricultural Products & Services	Chicago, Illinois	1957-03-04	0000007084	1902
ADBE	Adobe Inc.	Information Technology	Application Software	San Jose, California	1997-05-05	0000796343	1982

- שם החברה
- סימbol
- תחום החברה
- תת תחום
- מקום
- תאריך הצטרפות ל s&p500
- מפתח אינדקס החברה
- תאריך הקמת החברה

שמרנו את כל הדטא בקובץ csv לצורך
רישמה מסודרת של חברות שעלייהם המחק

לקחנו דатаה בעריה מהאתר - **glassdoor** קרא של שורה בטבלה

היא חברה וכל עמודה מייצגת:

The screenshot shows the Glassdoor website. At the top is a search bar with the placeholder "Search for job titles, companies, or key". Below the search bar are four navigation tabs: "Jobs" (with a briefcase icon), "Companies" (with a building icon, currently selected and highlighted in blue), "Salaries" (with a money bag icon), and "Careers" (with a leaf icon). The main content area features a large, vibrant photograph of a smiling woman with curly hair, wearing a tan jacket and a pink shirt, looking at her smartphone. The background of the photo is a city street with buildings. The entire screenshot is framed by a white border.

- שם החברה
- מטה התעשייה
- מקום
- מספר עובדים
- מספר ביקורות
- מספר משכורות
- מספר משרות
- דירוג כללי של דף החברה

באתר glassdoor השתמשנו ב-crawling ללקח את תגבות העובדים על חברות בהן הם מעסיקים, בין השנים 2017-2019. את כל התגבות שמרו בטבלה אחת חדשה.

כרגע שורה בטבלה מייצגת תגובה של עובד על חברתו שלו, וכל עמודה מייצגת לנו דירוג העובדים במגוון מאפיינים:

- **ציון הנהלה הבכירה**
- **ציון העסקה**
- **עובד לשעבר או עובד קיימים**
- **תאריך**
- **תפקיד**
- **יתרונות**
- **חרונות**
- **דירוג התגובה**
- **איזון בית- עבודה**
- **תרבות וערכים**
- **הזדמנויות קידום**
- **ציון סה"כ**
- **איזון עבודה/חיים**
- **ציון תרבות וערכים**
- **ציון גיון והכללה**
- **ציון ההזדמנויות קרירה**
- **ציון שכר והטבות**

Symbol	Security	GICS Sector	GICS Sub-Industry	Headquarters Location	Date added	CIK	Founded
AAPL	Apple Inc.	Information Technology	Technology Hardware, Storage & Peripherals	Cupertino, California	1982-11-30	0000320193	1977

>>>

לדוגמה הערך של החברה Apple בשני האתרים

glassdoor

Apple Location   

Jobs Companies Salaries Careers For Employers Post Jobs



 **Apple** Engaged Employer

Overview 38K Reviews 3.9K Jobs 101K Salaries 86 Q&A 11K Interviews 13K Benefits 11K Diversity [Follow](#) [+ Add a Review](#)

Apple Overview

Website: www.apple.com Headquarters: Cupertino, CA

Size: 10000+ Employees Founded: 1976

Type: Company - Public (AAPL) Industry: Computer Hardware Development

Apple Locations

Ang Mo Kio New Town (Singapore) 4.7 ★

Austin, TX 4 ★

Bukit Merah Estate (Singapore) 2 ★

Data cleaning



בנוסף לקלות הנתונים מזיקיפדיה ביצענו ניקוי ראשוני של Data: חברות שהופיעו בкопיות, או חברות שיש להם מעט תגבות כמו חברות אחזקה וכו'. לדוגמה חברת גול שמו פיעעה גם C-ALPHA כי היא הבעלים של החברה בתחלאס או חברות שהיא להם מעט מדי ביקורת בסיסglassdoor.

באמצעות API הוציאנו מידע פיננסי על כל חברת מניות משנת 2017 עד 2019 ושמרכנו בטבלה. כל

שורה הציגה החברה והעמודות הציגו:

The screenshot shows a financial analysis page for NVIDIA Corporation (NVDA). At the top, there's a search bar and a logo for 'Financial Modeling Prep'. Below that, the stock symbol 'NVDA' and price '\$408.22' are displayed, along with a red percentage '-2.95 (-0.72%)'. A navigation bar includes tabs for 'Financial Summary' (which is active), 'Financial Statements', 'Quarter Financials', 'Chart', and 'Fi...'. Under the summary tab, there's a rating section with five blue stars and a recommendation 'Strong Buy'. To the right is a 'Downloads' button with a cloud icon. The main table below contains various financial metrics:

Symbol	NVDA	DCF Unlevered	NVDA DCF ->
Price	\$408.22	DCF Levered	NVDA LDCF ->
Beta	1.751	ROE	18.66%
Volume Avg.	47.69M	ROA	10.61%
Market Cap	1.008T	Operating Margin	29.76% Strong Buy
Shares (2023)	2.51B	Debt / Equity	86.34%
52 Week Range	108.13-439.9	P/E	207.22 Strong Buy
1y Target Est	\$261.2857142857143	P/B	45.51 Strong Buy

- **תאריך**
- **דיווח**
- **שוויים לטווח קצר**
- **השקעות נטו**
- **חוויות החברה**
- **סכום הכל נכסים שותפים**
- **סה"כ בעלי המניות**
- **סה"כ הון עצמי**
- **סה"כ התחביבות ומחזיקי מניות**
- **השקעות**
- **ଓuid**

TobinsQ	Date	Company Name	Symbol
1.424687289	30-12-22	3M	MMM
2.162882683	31-12-21	3M	MMM
2.146057424	31-12-20	3M	MMM
2.281797829	31-12-19	3M	MMM
3.019009273	31-12-18	3M	MMM
3.701703359	29-12-17	3M	MMM
3.261547549	30-12-16	3M	MMM
2.7999246	31-12-15	3M	MMM
2.778771721	30-12-22	A. O. Smith	AOS
3.879819218	31-12-21	A. O. Smith	AOS
2.808037972	31-12-20	A. O. Smith	AOS
2.521769778	31-12-19	A. O. Smith	AOS
2.325073013	31-12-18	A. O. Smith	AOS
3.290124982	29-12-17	A. O. Smith	AOS
2.840698375	30-12-16	A. O. Smith	AOS
2.533243171	31-12-15	A. O. Smith	AOS
2.571629611	30-12-22	Abbott Laboratories	ABT
3.286841008	31-12-21	Abbott Laboratories	ABT
2.689710047	31-12-20	Abbott Laboratories	ABT
2.27032398	31-12-19	Abbott Laboratories	ABT
1.90523674	31-12-18	Abbott Laboratories	ABT
1.308464186	29-12-17	Abbott Laboratories	ABT
1.07784606	30-12-16	Abbott Laboratories	ABT
1.604094569	31-12-15	Abbott Laboratories	ABT
2.059523066	30-12-22	AbbVie	ABBV

אחריו שאספנו את הנתונים
על כל חברת, עשוינו
נירמו על שווי של
החברה ביחד עם הערכיהם
הפיננסיים שלה למשתנה
שנקרא Q tobins ולחרכיהם
בין 0 ל-18
עבור החברה tobins Q=
שווי החברה



尼克וי ומייזוג הנתוניים

- מיזגנו את כל הקבצים של הטבלאות לטבלה אחת מרוכזת .
- לאחר מכן ניקנו נתונים חוזרים בדעתם כמו כפליות של קטגוריות, ערכים נומריים ועוד...



EDA

**במקרים רבים כמות הנתונים שאנו
רוצים לעבד גדולה מאוד. אנחנו
זוקקים לדרר עיליה להבין כמות
גדולות של נתונים. וכך, באה לידי
ביטוי הויזואלייזציה.
גישה של ניתוח מערכי נתונים כדי
לסכם את המאפיינים העיקריים
שליהם,
לעתים קרובות משתמש בגרפיקה
סטטיסטית ובשיטות אחרות להדמיה
של נתונים." ~ויקיפדיה**

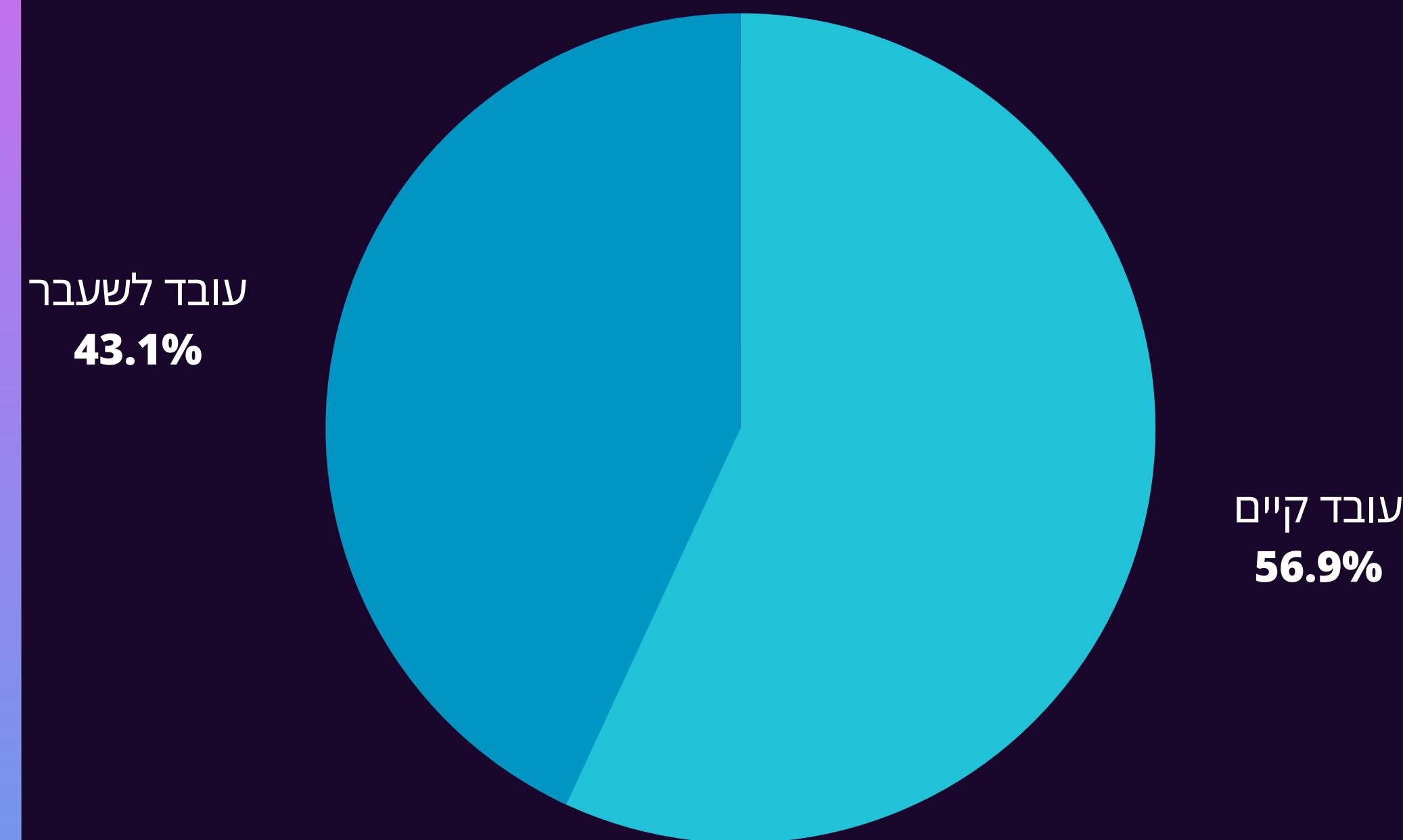




>>>

**בגרף זה ניתן לראות
שעובדים קיימים מבאים
יותר
מעובדים שפוטרו או
התפטרו.**

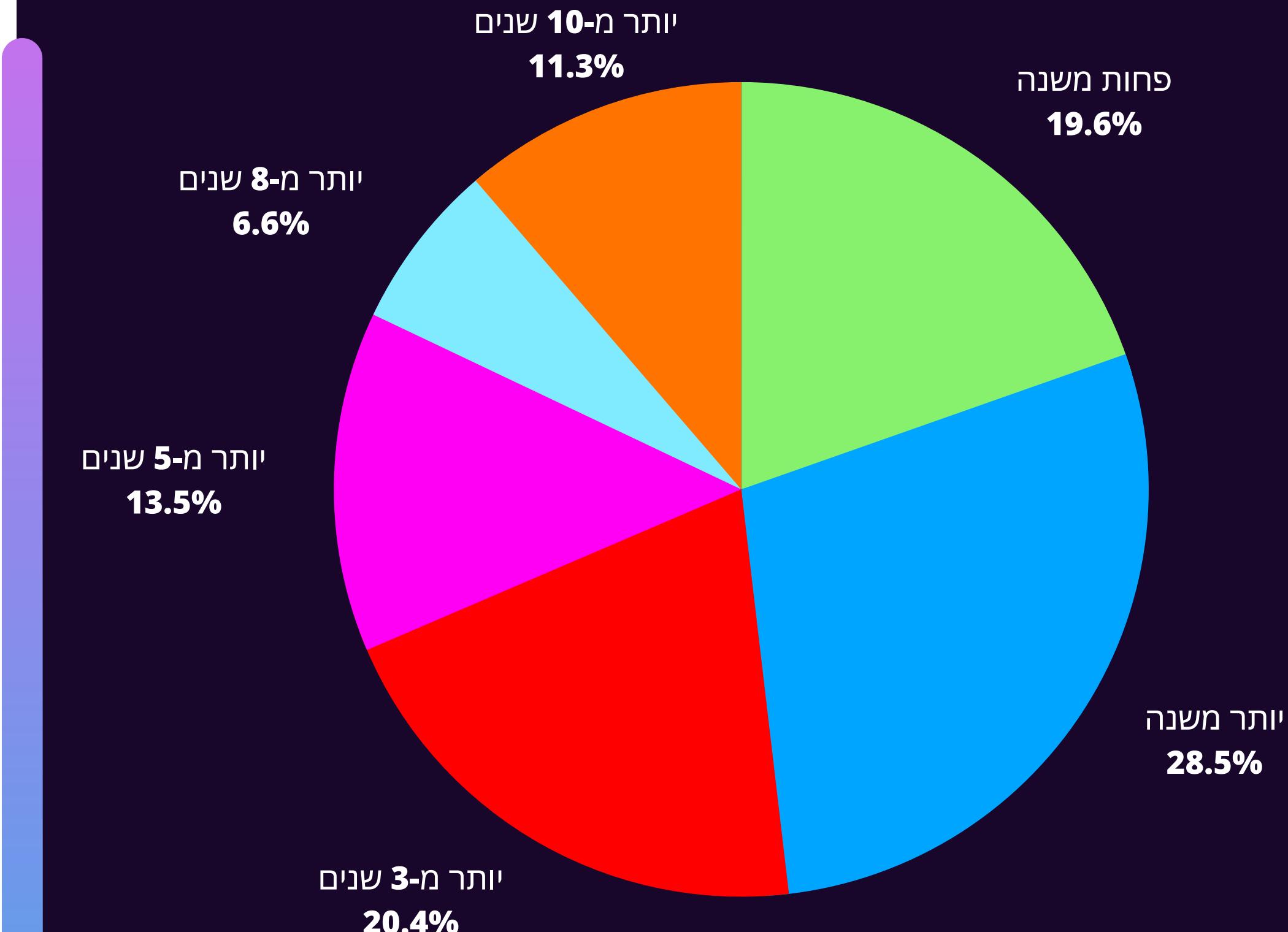
**סוג העובד שהגיע
על החברה**





**בגרף זה ניתן לראות את
התפלגות בין המגיבים
לבין זמן העסקתם
בחברה**

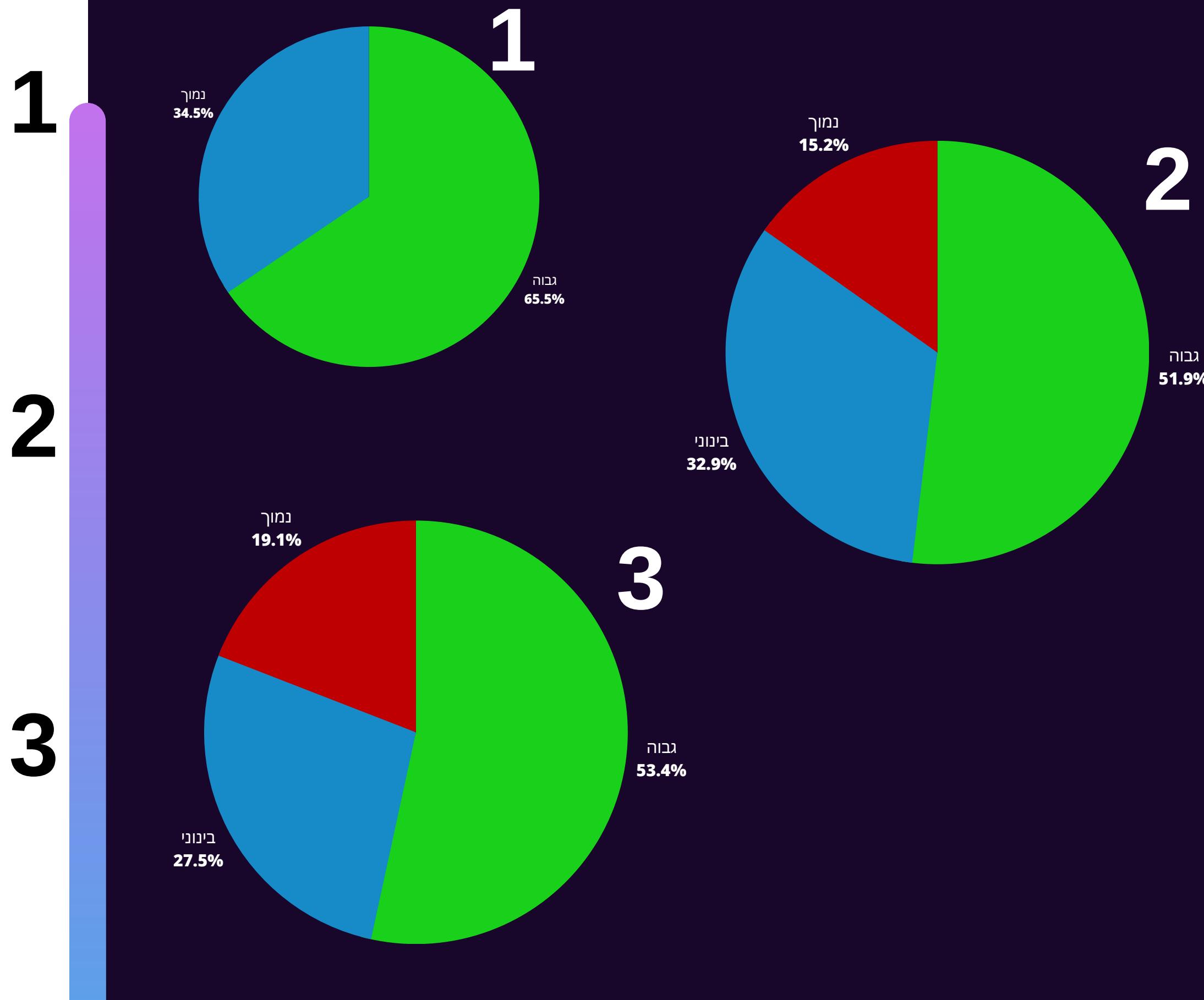
זמן העסקת העובד המבקר



דירוג העובדים לפי פרמטרים



האם העובד ממליץ על החברה

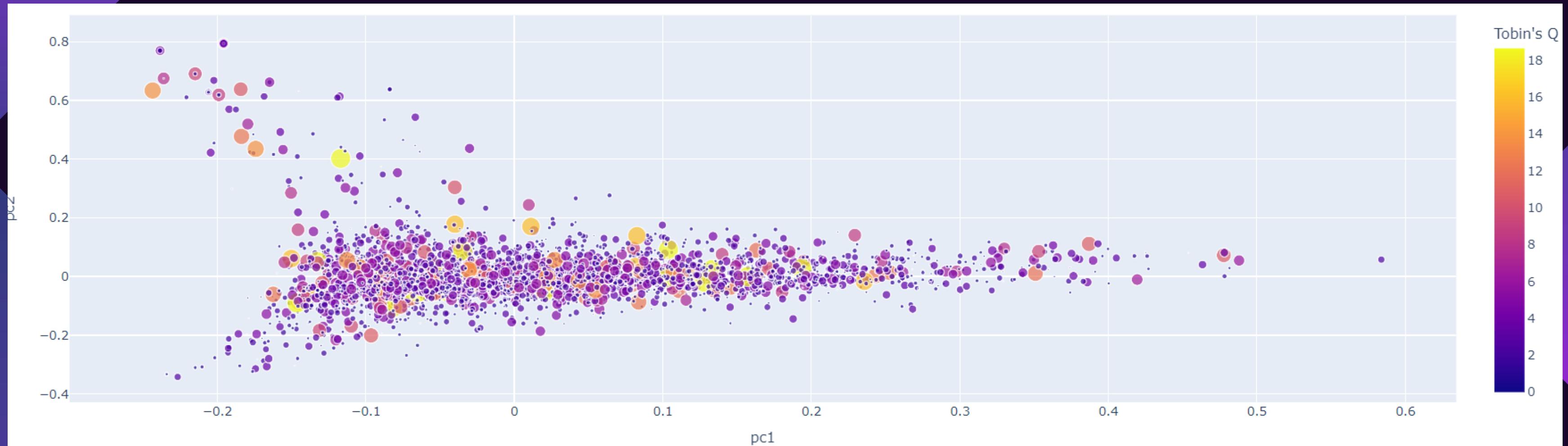


ציוו העובד להנהלת החברה

צפי העובד לעתיד החברה

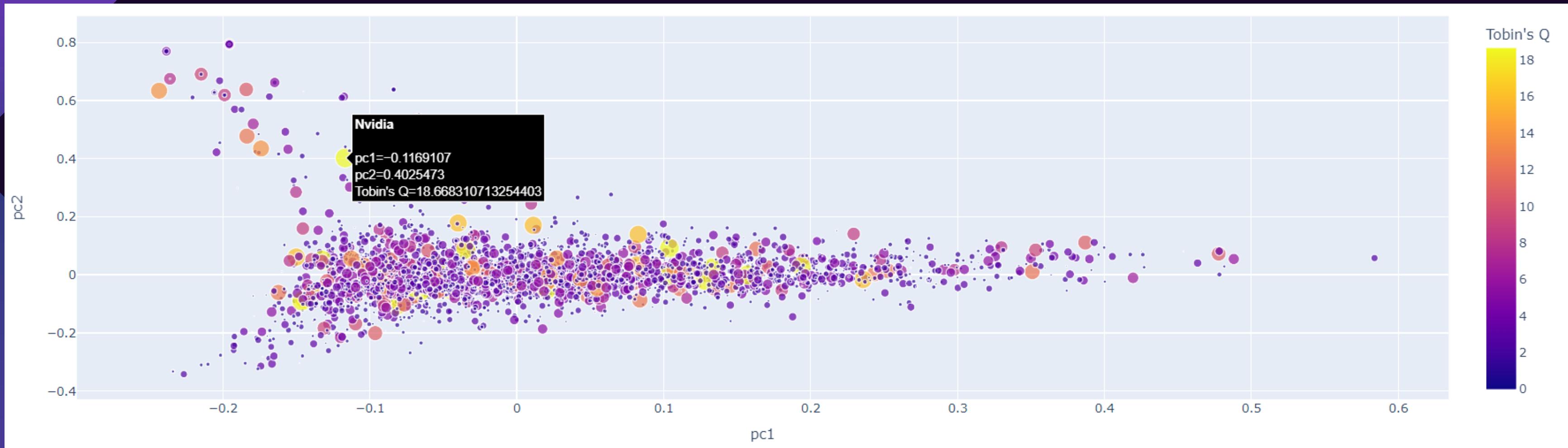


השתמשנו בPCA בכך לחריד את הממידים שנלקחו מעמודות הטקסט כדי שנוכל ליצור את כל התוצאות על גרפ אחד, בנוסף לכך ישו הרבה נתונים והמחשב קרס, לקחנו לכל חברה רק 10 תוצאות בצורה רנדומלית, בגרף שולפנו בוחנו את הקורולציה בין התוצאות שבחרנו לביצועי החברה.





לדוגמה ניתן לראות כאן את חברת Nvidia





Correlation matrix

מטריצת מותאם היא מטריצה מרובעת המציג את מקדמי המתחם בין שני משתנים. מקדמי מותאם מודדים כמה חזק ובאיזה כיוון שני משתנים מושרים בקשר ישר. מטריצת מותאם בוחנת לעיתים קרובות את הקשר בין משתנים שונים בניתו בסטטיסטיות רב-משתנים.

ניתן לראות כאן את הקרוולציה בין כל המאפיינים

	1	0.005131703	0.1573948	0.1625472	0.1546857	-0.2143967	-0.1286642	-0.1323042	-0.1871103	-0.08706481	-0.1834863	0.05960682	0.05187307	-0.02611261	0.01233116	1
Current/Former Employee	1	0.005131703	0.002017371	-0.02154915	-0.00209143	-0.01793486	-0.007574073	-0.008027962	-0.02619634	0.04086751	-0.03357276	0.03998216	0.01582126	-0.02907129	-0.05458054	0.8
Employment Length	0.005131703	1	0.3320684	0.3475449	-0.6660167	-0.4561813	-0.4555098	-0.546027	-0.4291421	-0.5717787	0.3339835	0.2419709	-0.07516737	0.0312617	0.6	
Recommend Level	0.1573948	0.002017371	1	0.3320684	0.3475449	-0.3501163	-0.2754412	-0.2760146	-0.30642	-0.2460022	-0.3669575	0.1226072	0.1067027	-0.08469116	-0.04356871	0.4
CEO Approval Level	0.1625472	-0.02154915	0.3320684	1	0.3458381	-0.3599944	-0.2524301	-0.2518488	-0.3288036	-0.2553566	-0.3492557	0.112899	0.1129303	-0.09387293	-0.02386588	0.2
Business Outlook Level	0.1546857	-0.00209143	0.3475449	0.3458381	1	-0.3599944	-0.2524301	-0.2518488	-0.3288036	-0.2553566	-0.3492557	0.112899	0.1129303	-0.09387293	-0.02386588	0
Total Score	-0.2143967	-0.01793486	-0.6660167	-0.3501163	-0.3599944	1	0.574584	0.5751654	0.6922528	0.5657699	0.7037263	-0.4105075	-0.3094381	0.1714141	-0.005165424	-0.2
Work/Life Balance Score	-0.1286642	-0.007574073	-0.4561813	-0.2754412	-0.2524301	0.574584	1	0.9120053	0.4673969	0.4322627	0.5527405	-0.2542031	-0.1884583	0.2020049	0.01928709	-0.4
Culture & Values Score	-0.1323042	-0.008027962	-0.4555098	-0.2760146	-0.2518488	0.5751654	0.9120053	1	0.4733019	0.4251449	0.5536568	-0.2513093	-0.1863875	0.2037234	0.02323963	-0.6
Career Opportunities Score	-0.1871103	-0.02619634	-0.546027	-0.30642	-0.3288036	0.6922528	0.4673969	0.4733019	1	0.5140003	0.6443536	-0.2955403	-0.2344557	0.1472482	-0.01007523	-0.8
Compensation and Benefits Score	-0.08706481	0.04086751	-0.4291421	-0.2460022	-0.2553566	0.5657699	0.4322627	0.4251449	0.5140003	1	0.4753283	-0.1684082	-0.1621125	0.1486359	-0.00261928	-0.1
Senior Management Score	-0.1834863	-0.03357276	-0.5717787	-0.3669575	-0.3492557	0.7037263	0.5527405	0.5536568	0.6443536	0.4753283	1	-0.3092989	-0.2378303	0.1731784	0.01540075	-0.3
Helpful Rate	0.05960682	0.03998216	0.3339835	0.1226072	0.112899	-0.4105075	-0.2542031	-0.2513093	-0.2955403	-0.1684082	-0.3092989	1	0.3951252	0.01373598	0.0346539	0.1
pc1	0.05187307	0.01582126	0.2419709	0.1067027	0.1129303	-0.3094381	-0.1884583	-0.1863875	-0.2344557	-0.1621125	-0.2378303	0.3951252	1	170.4297μ	0.006915041	0.1
pc2	-0.02611261	-0.02907129	-0.07516737	-0.08469116	-0.09387293	0.1714141	0.2020049	0.2037234	0.1472482	0.1486359	0.1731784	0.01373598	170.4297μ	1	-0.002956676	0.1
pc3	0.01233116	-0.05458054	0.0312617	-0.04356871	-0.02386588	-0.005165424	0.01928709	0.02323963	-0.01007523	-0.00261928	0.01540075	0.0346539	0.006915041	-0.002956676	0.1	1
	Current/Former Employee	Employment Length	Recommend Level	CEO Approval Level	Business Outlook Level	Total Score	Work/Life Balance Score	Culture & Values Score	Career Opportunities Score	Compensation and Benefits Score	Senior Management Score	Helpful Rate	pc1	pc2	pc3	



>>>

למידה מכונה

למוד למידת המוכנה השתמשנו בלמידה מונחת כר שעמודת המטרה שלנו היא: הערך הפיננסי של החברה

Tobin's Q





סוגי מודל

Random Forest

הוא אלגוריתם נפוץ של למידת מכונה, המשלב את הפלט של עצי החלטה מרובים כדי להגיע לתוצאה אחת. **קלות השימוש** והגמישות שלו היזנו את אימוצו, מכיוון שהוא מטפל בבעיות סיווג וגרסיה כאחד.

KNN

אלגוריתם השכן הקרוב או **k-Nearest Neighbors algorithm** הוא אלגוריתם חסר פרמטרים לסיווג ולגרסיה מקומית. בשני המקרים הקלט תלוי ב-**k** התצפיות הקרובות למרחב התכונות. NN-**k** יכול לשמש לסיווג או לגרסיה: NN-**k** לסיווג – בהינתן קלט של דוגמה חדשה, האלגוריתם משיכה לקבוצה. הדוגמה משוכנת למחלקה הנפוצה ביותר בקרב **k** השכנים הקרובים.

גרסיה לינארית

ניתוח גרסיה לינארית משמש לניבוי ערכו של משתנה על סמך ערכו של משתנה אחר. המשתנה שאתה רוצה לחזות נקרא המשתנה הבלתי תלוי. המשתנה שבו אתה משתמש כדי לחזות את הערך של המשתנה الآخر נקרא המשתנה הבלתי תלוי.



R2-תוצאות חיזוי המודל

```
[ ] train(LinearRegression(), features_df, target_df, Y_COLUMN)  
R2 score: 0.02408253371372182  
  
[ ] train(KNeighborsRegressor(), features_df, target_df, Y_COLUMN)  
R2 score: 0.2218384339722509  
  
▶ train(RandomForestRegressor(),features_df, target_df, Y_COLUMN)  
👤 R2 score: 0.8539848793916472
```

The best [R2 score](#) is **0.85** by the Random Forest model.

גרסיה לינארית
0.024

KNN
0.221

Random Forest
0.853





מסקנת המחקר

אחרי טיפול בדатаה לקחנו רק 10 תגבות לכל חברה ולפי זה אימנו את המודל. מכיוון שלא היה מספיק RAM במערכת המחשב קרס כמה פעמים. לא היה ניתן לחזות במדוקן את הקשר בין התגבות לערך הפיננסי של החברה.



**תודה
שצפתם!**