



# קלפי אספנות של ספורט אמריקאי

מוגש ע"י מרדכי אהרונסון, עבור קורס במדעי הנתונים במכון הטכנולוגי חולון

\$3,720,000



\$2,400,000

# הקדמה

## • הקדמה

- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- ויזואליזציה
- ניתוח נתונים
- מתקדם
- יישום והערכת
- ביצועים
- סיכום ומסקנות

## ▶ מדוע בחרתי בנושא של קלפי אספנות בתחום ספורט אמריקאי?

▶ לפני מספר שבועות צפיתי בסדרה ב- Netflix על בית מכירות פומביות בשם "גולדין" שמתמחה במכירות מוצרים אמריקאים מובהקים, במיוחד בתחום ספורט, קומיקס וכל שאר, התאהבתי בסדרה וברעיון של אספנות קלפי ספורט, להפתעתי, מדובר בתחום מאוד מושקע שמגלגל מיליארדים דולרים, סכומים עצומים.

▶ לדוגמה, בסדרה היה חיפוש אחרי קלף יחיד במינו שעדיין לא נגלה מחפיסת קלפים סגורה (ליתר הדיוק, קופסה יוקרתית של חברת Panini בשם Flawless שעולה בין \$8000 ל-\$25000, המחיר מאמיר שלא מוצאים את הקלף היוקרתי והיחיד במינו), הקלף הוא של השחקן לברון ג'יימס, שמכיל בתוכו 3 טלאים מגופיות משחק שלו, ממשחקים שונים, הקלף נמכר בסכום של 2.4 מיליון דולר, בית מכירות פומביות "גולדין" זה שביצע את המכירה (כמתווך בין מוצא/מוכר הקלף למציעים במכירה פומביות), קלף נוסף למשל הוא על שחקן בייסבול בשם הונוס ונגר, מדובר בקלף משנת 1909-1911 שלפני הערכות נותרו רק כ-50-60 קלפים כאלה וגם אם מצבם לא במצב מינט, ניתן למכור אותם בסכומי עתק, למשל, בגולדין הוא נמכר בסכום של 3,720,000 דולר.

▶ כך סיקרן אותי מהם מאפיינים שיכולים לעשות את הקלף ליקר – שזה למעשה שאלת המחקר שלי



# מקורות הנתונים והרכשה

▶ מקור העיקרי של נתונים להרכשה

▶ הוא כמובן, האתר החביב עלי, בית מכירות פומביות "גולדין" שמכיל בתוכו עשרות אלפים מכירות שנסגרו כבר.

The screenshot displays the Goldin's auction website interface. At the top, there are tabs for "Results", "View Live Items", and "View Past Auctions". Below this is a search bar and filters for "Auction" and "Marketplace". The main content area shows a grid of auction items, each with a thumbnail image, a title, a price, and the number of bids. The items include a 1909-11 T206 White Border Honus Wagner - PSA FR 1.5, a Michael Jordan 1992 Olympic "Dream Team" Game-Used, Photo-Matched, Signed, Inscribed Jersey, a 2003-04 Upper Deck Exquisite Collection Rookie Patch Autograph (RPA) #78 LeBron James Signed, a 2020-21 Panini Flawless Triple Logoman #3LC-1B.1 LeBron James Parallel (RPA) #1/1 - PSA Authentic, a 2003-04 Upper Deck Exquisite Platinum NFL Shield Autograph Parallel (RPA) #78 LeBron James Signed, and a 2020 Panini National Treasures Herbert Signed NFL Shield Rookie.

Item	Price	Bids	Ended
1909-11 T206 White Border Honus Wagner - PSA FR 1.5	\$3,720,000	27 bids	Mon, 10/10/22, 5:30 AM
Michael Jordan 1992 Olympic "Dream Team" Game-Used, Photo-Matched, Signed, Inscribed Jersey	\$3,030,000	40 bids	Thu, 5/25/23, 7:46 AM
2003-04 Upper Deck Exquisite Collection Rookie Patch Autograph (RPA) #78 LeBron James Signed	\$2,400,000	24 bids	Sun, 10/23/22, 5:30 AM
2020-21 Panini Flawless Triple Logoman #3LC-1B.1 LeBron James Parallel (RPA) #1/1 - PSA Authentic	\$2,400,000	25 bids	Sun, 6/26/22, 5:49 AM
2003-04 Upper Deck Exquisite Platinum NFL Shield Autograph Parallel (RPA) #78 LeBron James Signed	\$2,100,000	24 bids	Sun, 6/26/22, 5:30 AM
2020 Panini National Treasures Herbert Signed NFL Shield Rookie	\$1,800,000	31 bids	Sun, 9/11/22, 7:02 AM

• הקדמה

• מקורות הנתונים

והרכשה

• ניתוח ראשוני וטיוב

• ויזואליזציה

• ניתוח נתונים

מתקדם

• יישום והערכת

ביצועים

• סיכום ומסקנות

# מקורות הנתונים והרכשה

## ▶ מקורות מידע נוספות הן:

▶ מקורות מידע שמכילות רשימות קצרות, שלא הייתי צריך לכתוב קוד במיוחד עבורן בגלל שהן מכילות לכל היותר 15 רשומות, לדוגמה:

- מיהם 15 שחקני בייסבול הטובים ביותר בכל הזמנים: בייב רות', ווילי מייס, האנק אהרון, ועוד.. (מידע מאתר ESPN)
- מהן חברות שמייצרות את קלפי אספנות בתחום ספורט: Upper Deck, Panini, SkyBox.. (מידע מויקיפדיה ומאתרים נוספים)
- אילו צבעים שמייחדים את הקלפים (דרגות נדירות של קלפים). (מידע מאתר Panini)

- הקדמה

- **מקורות הנתונים**

- והרכשה**

- ניתוח ראשוני וטיוב

- ויזואליזציה

- ניתוח נתונים

- מתקדם

- יישום והערכת

- ביצועים

- סיכום ומסקנות

# מקורות הנתונים והרכשה

## ▶ מכשולים בהרכשת נתונים מאתר גולדין:

▶ היו קשיים טכניים בהרכשת נתונים מאתר גולדין כמפורט להלן:

- בדף פנימי של קלף כלשהו היו תגיות מרובות בעלות Class זהים, ומבנה היררכי של תגיות שלא איפשר לשאוב נתונים באופן יעיל, התגברתי על כך בכך ששאבתי נתונים מדף תוצאות של חיפוש ולא בדפים פנימיים כמו מספר bids או שעת סגירה של מכירה פומבית, גם ככה, לא הצלחתי לשאוב באופן מושלם את הנתונים, היו לא מעט נתונים שלא הצלחתי לשאוב אבל הגעתי לנפח נתונים שלדעתי מספיק.
- לאחר ששאבתי את כל הנתונים, היה חסר מידע שהוא description (תיאור המוצר), ניסיתי לשאוב את המידע מדפים פנימיים וזה דווקא הצליח כי לתגית של description היה id ייחודי, הבעיה היא שלמעט מהמוצרים היה description רציני ולשאר ללא description, בסופו של דבר, במהלך ניתוח נתונים ויתרתי על המידע מdescription.

▶ כמות נתונים שקיבלתי הוא: 24240 רשומות  $\times$  10 עמודות = 242400 נתונים

• הקדמה

• מקורות הנתונים

והרכשה

• ניתוח ראשוני וטיוב

• ויזואליזציה

• ניתוח נתונים

מתקדם

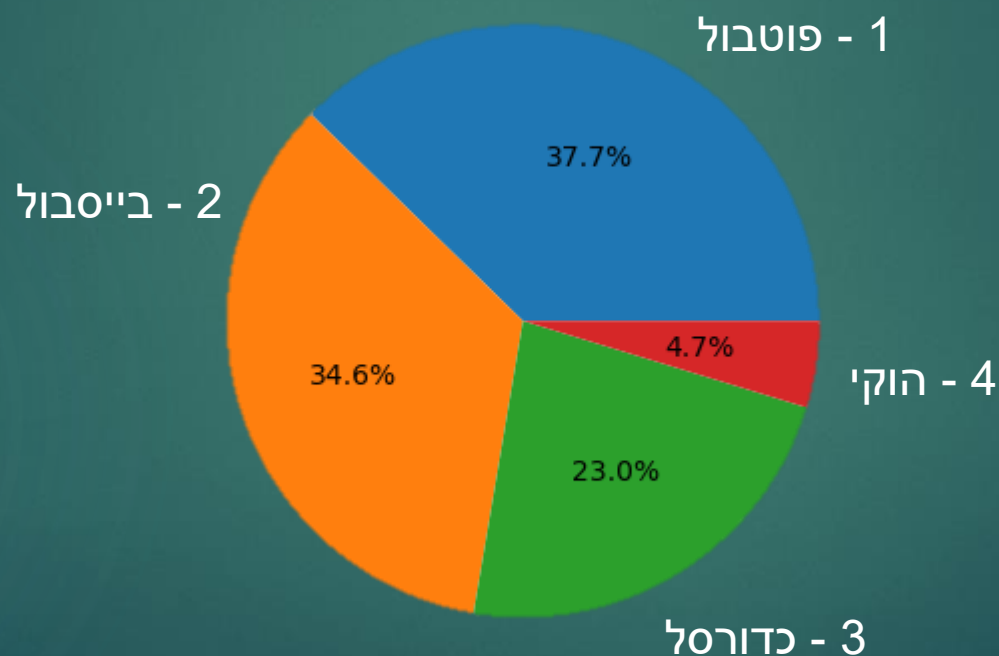
• יישום והערכת

ביצועים

• סיכום ומסקנות

# ניתוח ראשוני וטיוב

▶ לאחר שביצעתי הרכשת נתונים, היו לי 4 Data Frames, שכל אחד היה בתחום ספורט מסוים: פוטבול, בייסבול, כדורסל והוקי, לפני שאיחדתי אותם ל-DataFrame אחד, נתתי לכל אחד מספר קטגורי כך ש:



1. 1 - פוטבול

2. 2 - בייסבול

3. 3 - כדורסל

4. 4 - הוקי

▶ כאן תרשים בצורת פאי שמראה אופן חלוקת קלפים לתחומי ספורט שונים.

- הקדמה
- מקורות הנתונים
- והרכשה
- **ניתוח ראשוני וטיוב**
- ויזואליזציה
- ניתוח נתונים
- מתקדם
- יישום והערכת
- ביצועים
- סיכום ומסקנות

# ניתוח ראשוני וטיוב

- **הקדמה**
  - **מקורות הנתונים**
  - **והרכשה**
  - **ניתוח ראשוני וטיוב**
  - **ויזואליזציה**
  - **ניתוח נתונים**
  - **מתקדם**
  - **יישום והערכת**
  - **ביצועים**
  - **סיכום ומסקנות**
- ▶ רוב הנתונים שכן הצלחתי לשאוב הגיעו בשלמותם, מה שהגיע באופן חלקי, הסרתי אותם (כמו כמות bids), כמו כן, הסרת כפילויות בגלל שנאלצתי לבצע את שאיבת נתונים שוב ושוב כך שקיבלתי נתונים כפולים. לגבי תיקון חריגים, לא היו נתונים חריגים, רק נתון אחד שנובע מטעות הקלדה – שנת קלף 22020 במקום 2020 וריכוז של מחירים שהם מעל \$200000 היה דליל והחלטתי לוותר עליהם (הם בעצם היו תקינים אבל היה מוטב לסלק אותם מ Data Frame בגלל שהם הפריעו לתצוגת גרפים).
- ▶ הנתונים ב description היו מאוד חלקים, למעט רשומות היה description ולשאר היה ריק או מידע סתמי כמו "Please note that the population report information cited in a product description is accurate at the time a lot is posted for auction and is subject to change." ויתרתי על שימוש בנתונים ב description ובמקום זאת השתמשתי בניתוח טקסט על כותרת של מוצר, מה שהתגלה כמאוד מועיל.

# ניתוח ראשוני וטיוב

השתמשתי בכלי CountVectorizer כדי למצוא מהן 100 מילים נפוצות ביותר בכותרת של קלפי ספורט ולהלן הרשימה:

'04', '06', '08', '10', '100', '13', '15', '18', '19', '1996', '1997', '1998', '1999', '20', '2000', '2003', '2007', '2017', '2018', '2019', '2020', '2021', '2022', '21', '22', '23', '25', '50', '97', '98', '99', 'Authentic', 'Autograph', 'Autographs', 'BGS', 'Beckett', 'Black', 'Blue', 'Bowman', 'Bryant', 'Card', 'Chrome', 'Collection', 'DNA', 'Deck', 'Donruss', 'Draft', 'Dual', 'EX', 'Encased', 'Exquisite', 'Finest', 'Fleer', 'GEM', 'Game', 'Gold', 'Green', 'James', 'Jersey', 'Joe', 'Jordan', 'Jr', 'Kobe', 'LeBron', 'MINT', 'MT', 'Metal', 'Michael', 'Mosaic', 'NM', 'National', 'Optic', 'Orange', 'PSA', 'Panini', 'Patch', 'Picks', 'Pop', 'Premium', 'Prizm', 'Prospect', 'Prospects', 'Purple', 'Red', 'Refractor', 'Relic', 'Rookie', 'SGC', 'SP', 'Select', 'Signatures', 'Signed', 'Silver', 'SkyBox', 'The', 'Tom', 'Topps', 'Treasures', 'Upper', 'VG'

• הקדמה

• מקורות הנתונים

והרכשה

• ניתוח ראשוני וטיוב

• ויזואליזציה

• ניתוח נתונים

מתקדם

• יישום והערכת

ביצועים

• סיכום ומסקנות



# ניתוח ראשוני וטיוב

## ומכאן סיווגתי באופן ידני את הרשימה לקטגוריות (סיננתי כמה מילים ולא הכנסתי לקטגוריות):

- 04,06,100,21... - יכול לייצג כל מספר כמו מספר גופיה או שנה או כמות קלפים שהונפקו
- 2003,2020,2021,2022 – מייצג את שנת ייצור
- Authentic, Autograph, Autographs, Signatures, Signed – תוספות מיוחדות לקלף כמו קלף חתום או קלף עם פגם בייצור
- BGS, PSA, Beckett, GEM, DNA, SGC – חברות שנותנות דירוג למצב קלף
- Black, Blue, Chrome, Gold – צבעים של קלפים, שמייצגים לרוב רמת נדירות של קלף
- Optic, Mosaic, Prizm, Flawless – מייצגים סדרות קלפים, יש סדרות יוקרתיות וישנן פחות, בעיקר של חברת Panini
- James, Kobe, Jordan, Bryant – שמות של שחקנים
- MINT, MT, NM, EX, VG – מצב קלף
- Panini, Topps, SkyBox, Bowman – חברות שמייצרות קלפים

• הקדמה

• מקורות הנתונים

והרכשה

• ניתוח ראשוני וטיוב

• ויזואליזציה

• ניתוח נתונים

מתקדם

• יישום והערכת

ביצועים

• סיכום ומסקנות

# ויזואליזציה

- הקדמה
  - מקורות הנתונים
  - והרכשה
  - ניתוח ראשוני וטיוב
  - **ויזואליזציה**
  - ניתוח נתונים
  - מתקדם
  - יישום והערכת
  - ביצועים
  - סיכום ומסקנות
- ▶ יצרתי מספר לא קטן של תרשימים כדי לבחון את המאפיינים, השארתי את חלוקת נתונים לפי ענף ספורט על כנה, כדי שאוכל לבחון בנפרד מהם מאפיינים בולטים עבור כל ענף ספורט.
- ▶ בדקתי את הנתונים לפי מה שסיווגתי לקטגוריות:
- שמות שחקנים
  - תוספות מיוחדות (האם כרטיס חתום)
  - חברות דירוג של קלף
  - צבעים של קלפים (רמת נדירות)
  - סוג חבילה של קלפים
  - מצב קלף
  - חברות שמייצרות את קלפים
- ▶ ונתחיל עם שחקנים...

# ויזואליזציה

► חיפשתי בגוגל עבור כל ענף ספורט מהם 15 שחקנים הכי טובים בכל זמנים ולהלן הרשימה לדוגמה, עבור כדורסל:



- הקדמה
- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- ויזואליזציה
- ניתוח נתונים
- מתקדם
- יישום והערכת
- ביצועים
- סיכום ומסקנות

# ויזואליזציה

► וכאן לדוגמה עבור שחקני בייסבול בכל הזמנים:



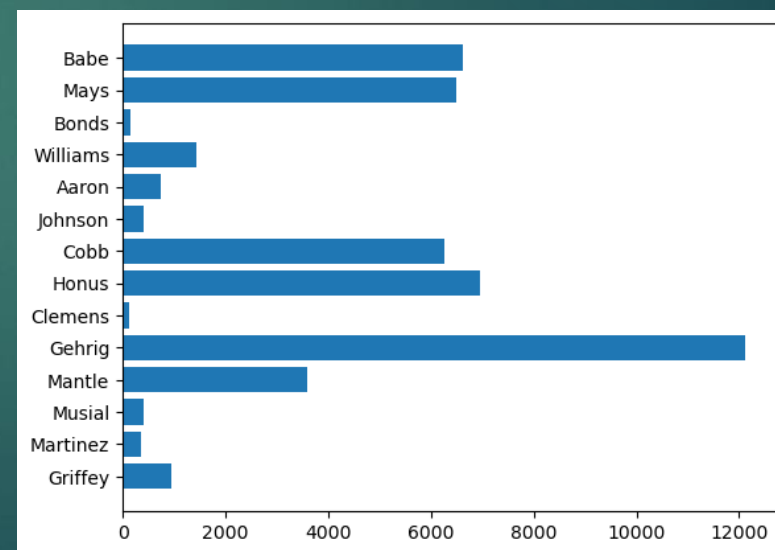
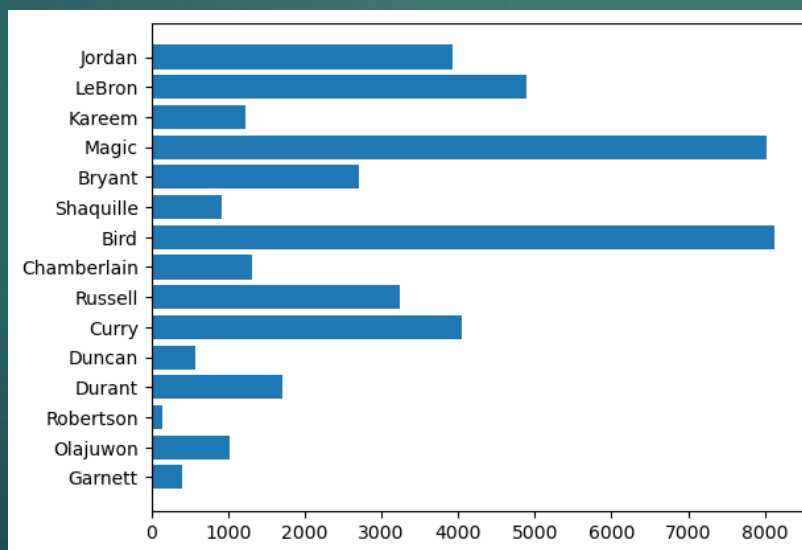
- הקדמה
- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- **ויזואליזציה**
- ניתוח נתונים
- מתקדם
- יישום והערכת
- ביצועים
- סיכום ומסקנות



# ויזואליזציה

► וכאן תרשימים עבור שמות של שחקנים בכדורסל ובבייסבול לדוגמה:

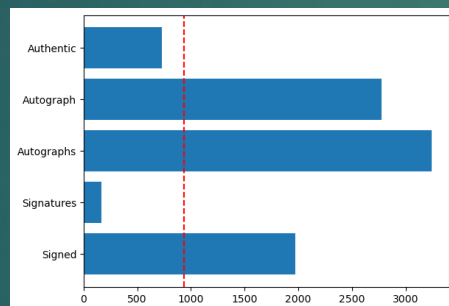
► ניתן לראות שאם שחקן הכי טוב בכדורסל הוא מייקל ג'ורדן אבל קלפים שלו לא זוכים לממוצע גדול ביותר, ממוצע של קלפים של לארי בירד ומג'יק ג'ונסון זוכה להיות משמעותי, כמו כן, בייב רות' שנחשב לגדול ביותר שבשחקני בייסבול אבל ממוצע סכום של קלפים של גריג זוכה ליותר, באופן משמעותי, אז איך בדקתי את זה? השתמשתי בRegular Expressions כדי לבדוק אם שם של שחקן מופיע בכותרת ולפניכם גרפים:



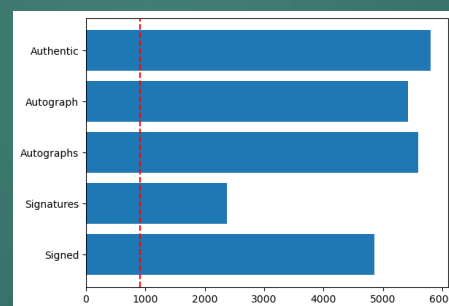
- הקדמה
- מקורות הנתונים והרכשה
- ניתוח ראשוני וטיוב
- ויזואליזציה
- ניתוח נתונים מתקדם
- יישום והערכת ביצועים
- סיכום ומסקנות

# ויזואליזציה

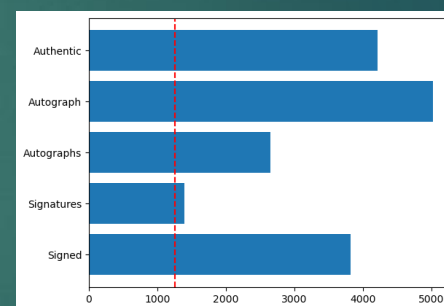
► וכאן בדקתי אם חתימה על הקלף מעניקה ערך יותר גדול? איך אני עונה על השאלה? ובכן, לקחתי את ממוצע סכום כל הקלפים עבור כל ענף ושאלתי אם קלף חתום עולה על הממוצע ולפניכם התוצאות:



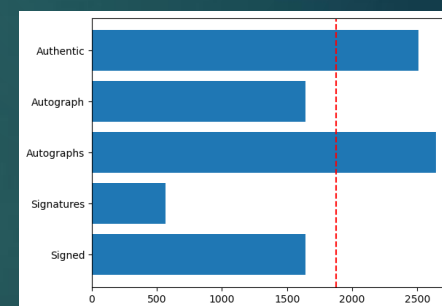
הוקי



כדורסל



פוטבול



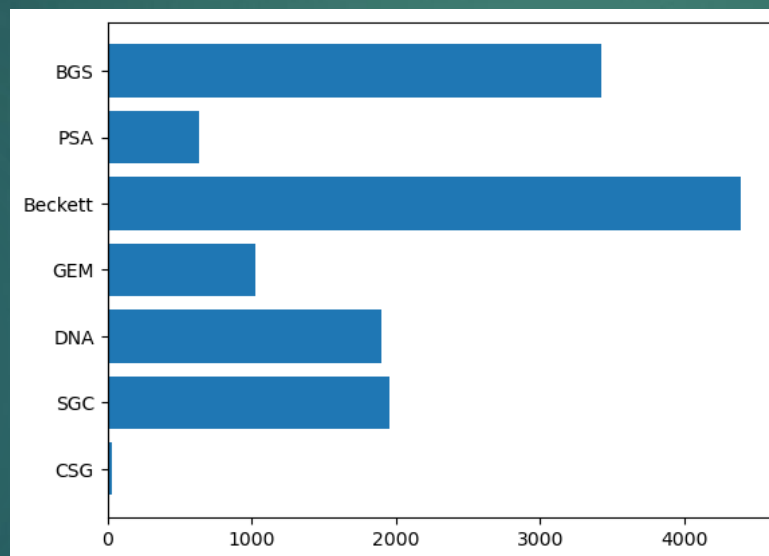
בייסבול

• כאן ניתן לראות שברוב המקרים קלף חתום או אותנטי עולה על הממוצע

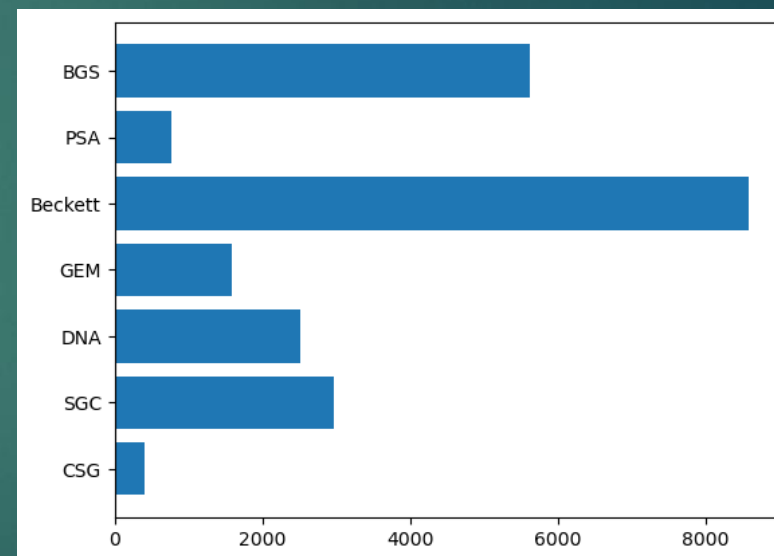
- הקדמה
- מקורות הנתונים והרכשה
- ניתוח ראשוני וטיוב
- ויזואליזציה
- ניתוח נתונים מתקדם
- יישום והערכת ביצועים
- סיכום ומסקנות

# ויזואליזציה

► ומה לגבי חברות דירוג מסוימת שמדרגת קלף, מעניקה את המשקל המשמעותי של המחיר של הקלף, הבה נבדוק עם תרשימים:



בייסבול

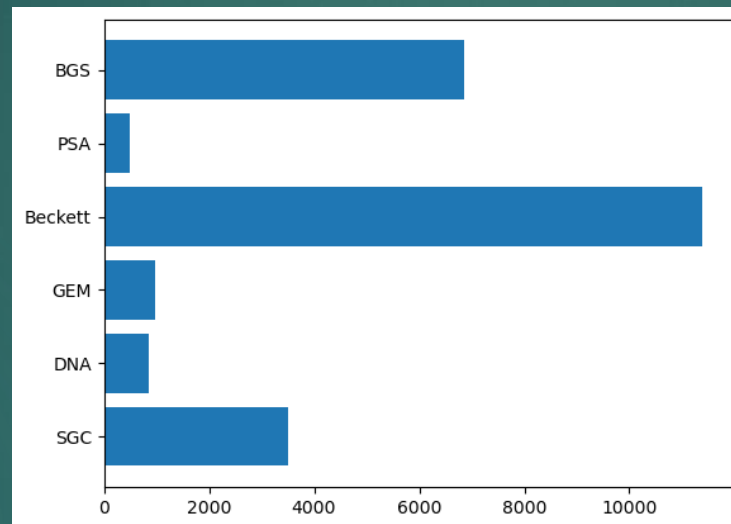


פוטבול

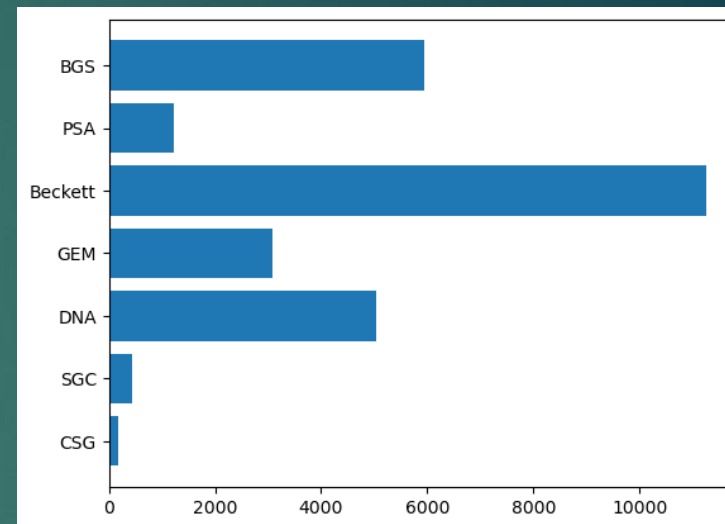
- הקדמה
- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- **ויזואליזציה**
- ניתוח נתונים
- מתקדם
- יישום והערכת
- ביצועים
- סיכום ומסקנות

# ויזואליזציה

- הקדמה
- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- **ויזואליזציה**
- ניתוח נתונים
- מתקדם
- יישום והערכת
- ביצועים
- סיכום ומסקנות



הוקי



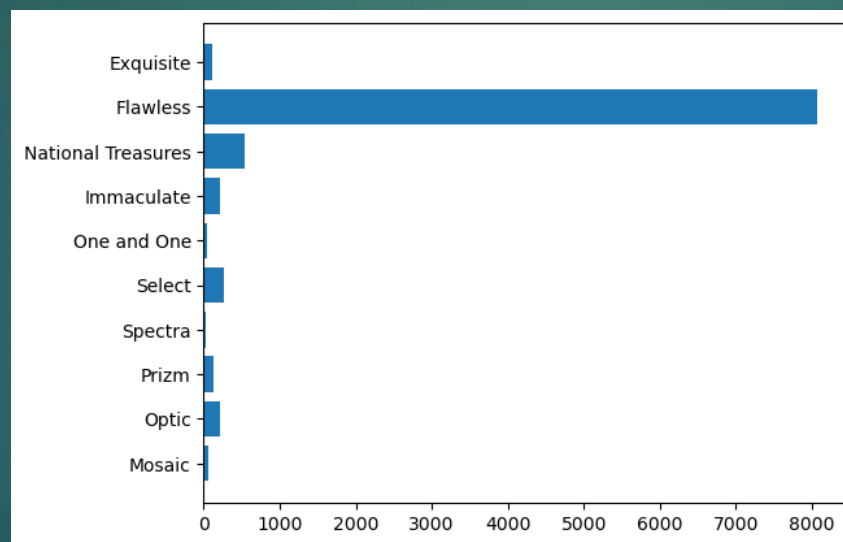
כדורסל

- כאן ניתן לראות באופן מובהק, עבור כל ענפי ספורט שחברת דירוג (BGS) Beckett זה שם נוסף של אותה החברה) שמעניקה דירוגים נותנת משקל מאוד משמעותי לערך הקלף

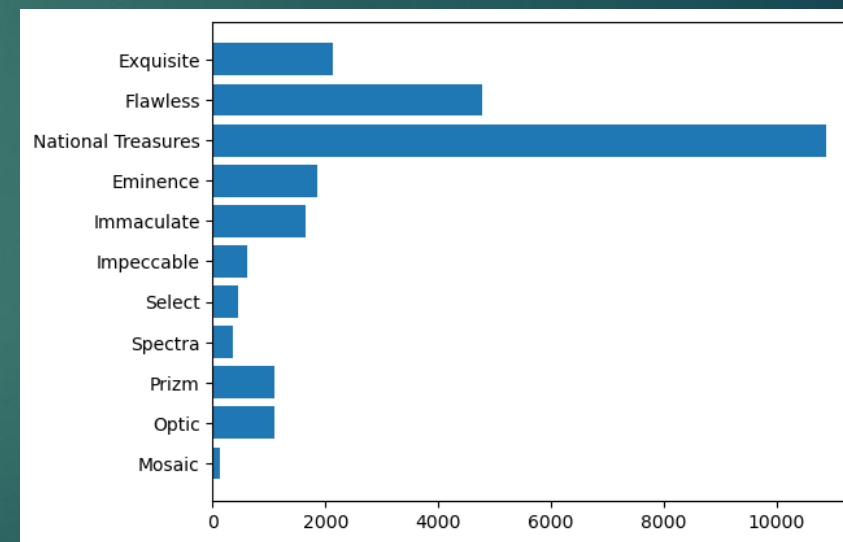


# ויזואליזציה

▶ אם סוגים של חפיסות מעניקים ערך מחיר יותר גדול לקלפים? לדוגמה, חפיסת קלפים Flawless (שמכילה רק 10 קלפים) שעולה בסביבות \$25000-\$8000 מעניקה ערך לקלפים שלה?



בייסבול

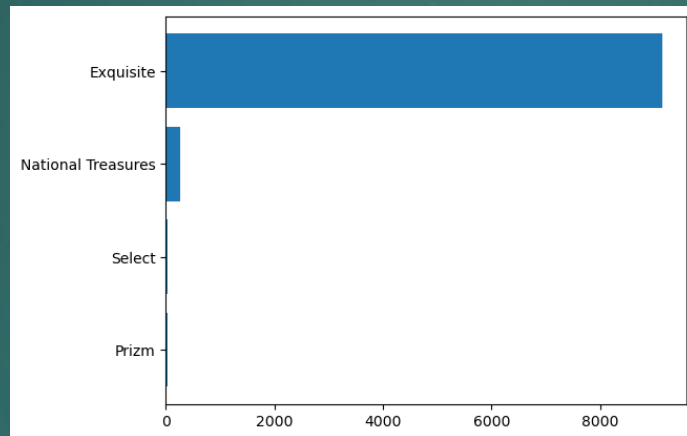


פוטבול

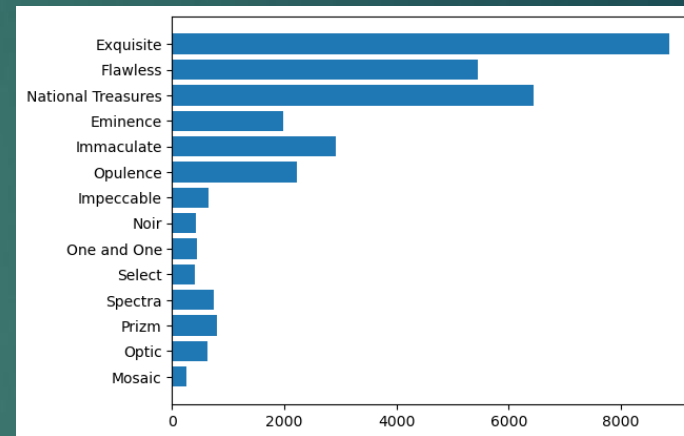
- הקדמה
- מקורות הנתונים והרכשה
- ניתוח ראשוני וטיוב
- ויזואליזציה
- ניתוח נתונים מתקדם
- יישום והערכת ביצועים
- סיכום ומסקנות

# ויזואליזציה

- הקדמה
- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- **ויזואליזציה**
- ניתוח נתונים
- מתקדם
- יישום והערכת
- ביצועים
- סיכום ומסקנות



הוקי



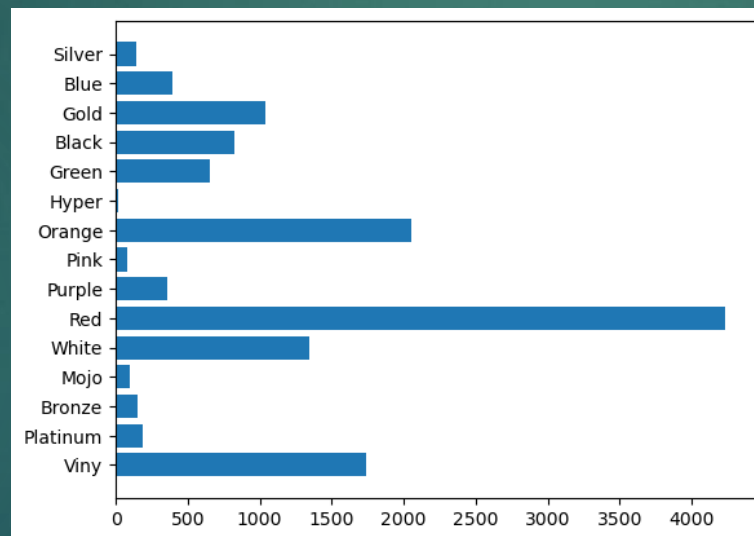
כדורסל

- כאן ניתן לראות שFlawless תורם חלק משמעותי אבל רק חלק מענפי ספורט, כאן ניתן לראות שגם National Treasures ו- Exquisite משחקים תפקיד משמעותי

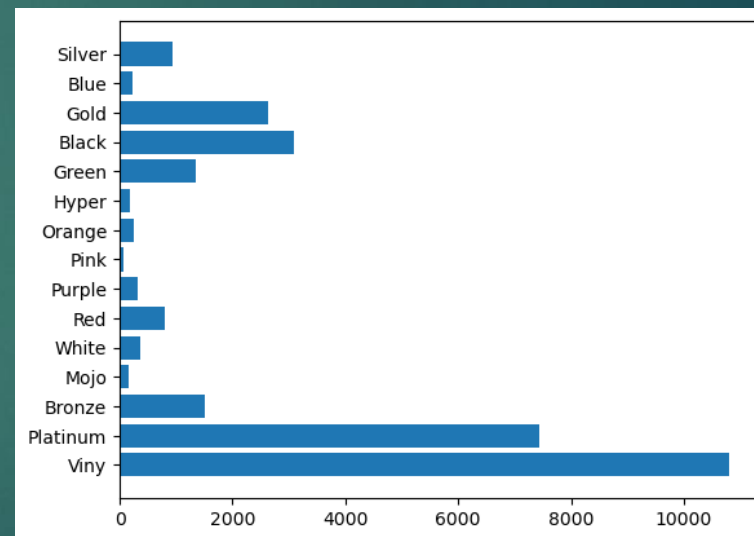
# ויזואליזציה

- הקדמה
- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- **ויזואליזציה**
- ניתוח נתונים
- מתקדם
- יישום והערכת
- ביצועים
- סיכום ומסקנות

▶ מה לגבי צבעים של קלפים? אם דרגות נדירות של קלפים משחקים תפקיד במחיר מכירה סופית? הבה נבדוק:



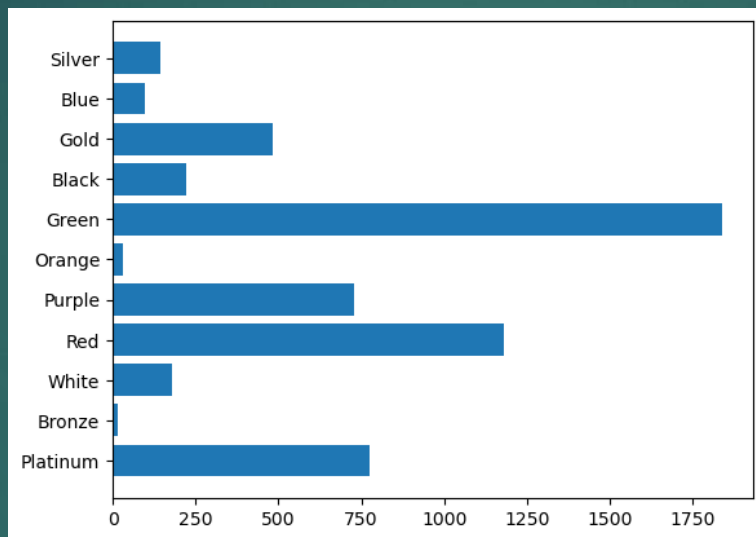
בייסבול



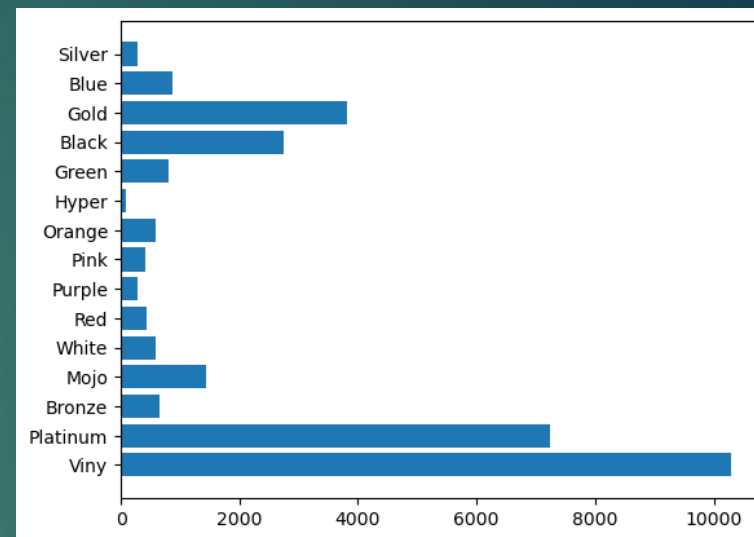
פוטבול

# ויזואליזציה

- הקדמה
- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- ויזואליזציה
- ניתוח נתונים
- מתקדם
- יישום והערכת ביצועים
- סיכום ומסקנות



הוקי



כדורסל

- התוצאות לא מאוד חד-משמעיות בגלל שRed וGreen בולטים בענפי ספורט מסוימים, להערכתך, מדובר בשמות של קבוצות כמו 'Boston Red Sox' או שם משפחה Green אבל מה שכן, Viny ו-Platinum די בולטים.

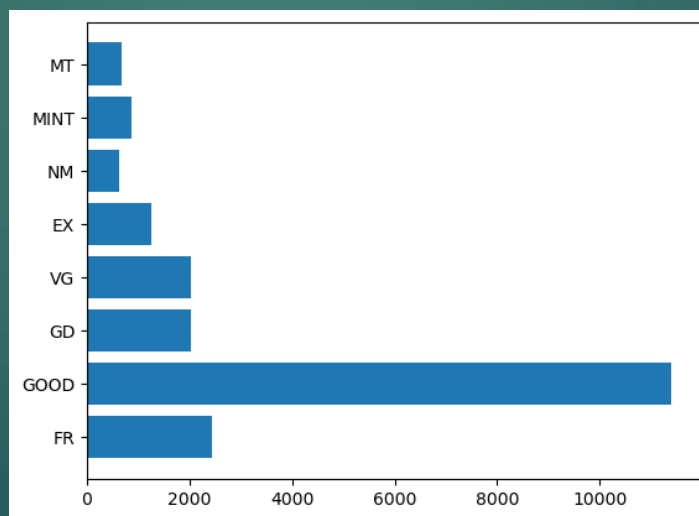


# ויזואליזציה

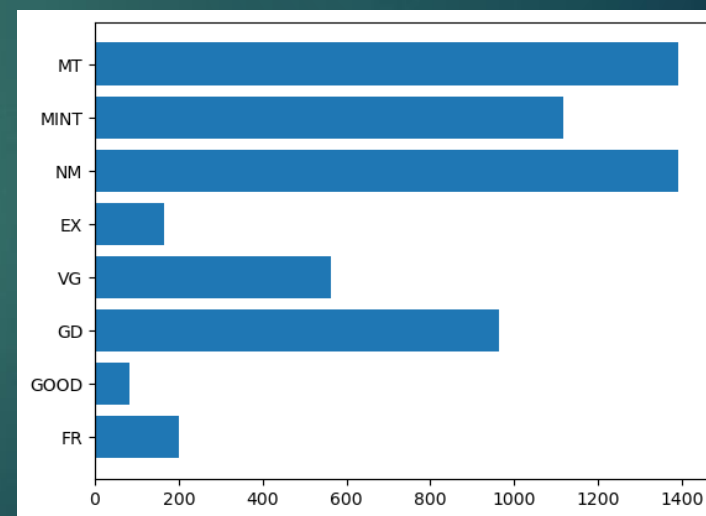
- הקדמה
- מקורות הנתונים והרכשה
- ניתוח ראשוני וטיוב
- **ויזואליזציה**
- ניתוח נתונים מתקדם
- יישום והערכת ביצועים
- סיכום ומסקנות

▶ אם מצב הקלף מעניק לקלף את הערך? המצב הטוב ביותר נקרא MINT או MT (גם GEM אבל לא הכנסתי לכאן את המילה כי הוא לרוב בא עם צמד מילים GEM MINT) ו FR (Fair) הוא "סביר" ונחשב לרמת איכות נמוכה ויש עוד מצבי ביניים, בוא נבדוק אם יש להם השפעה על המחיר:

איך ייתכן?! קלפים במצב FR או GOOD בביסבול יותר יקרים?!  
אולי בגלל שלרוב מדובר בקלפים ישנים יותר ולכן נדירים יותר ומכאן ערכם עולה



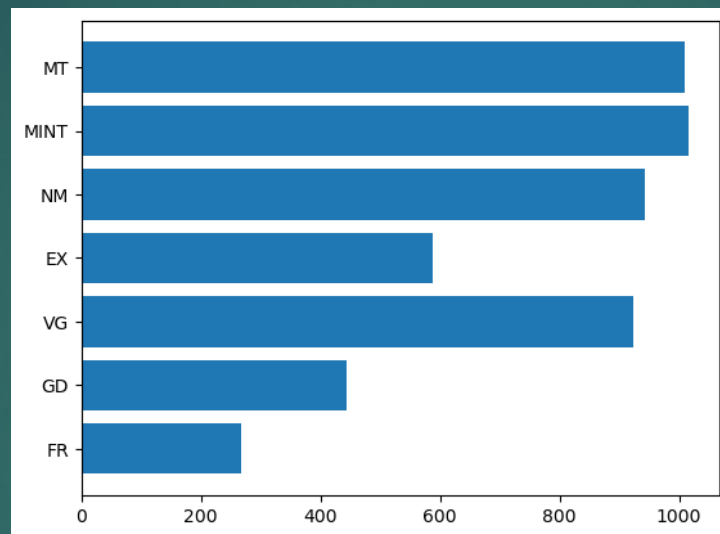
ביסבול



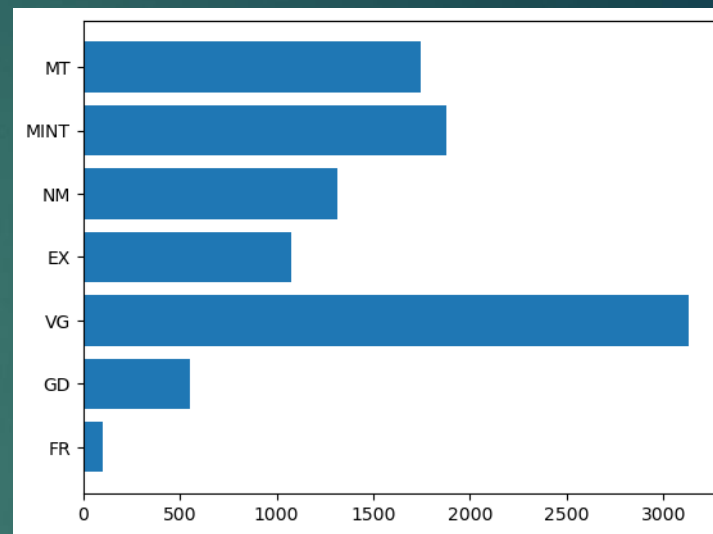
פוטבול

# ויזואליזציה

- הקדמה
- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- **ויזואליזציה**
- ניתוח נתונים
- מתקדם
- יישום והערכת
- ביצועים
- סיכום ומסקנות



הוקי

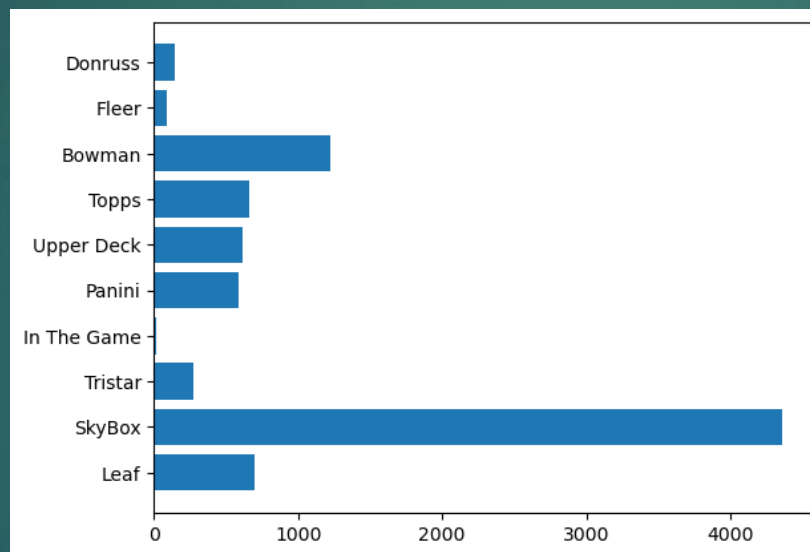


כדורסל

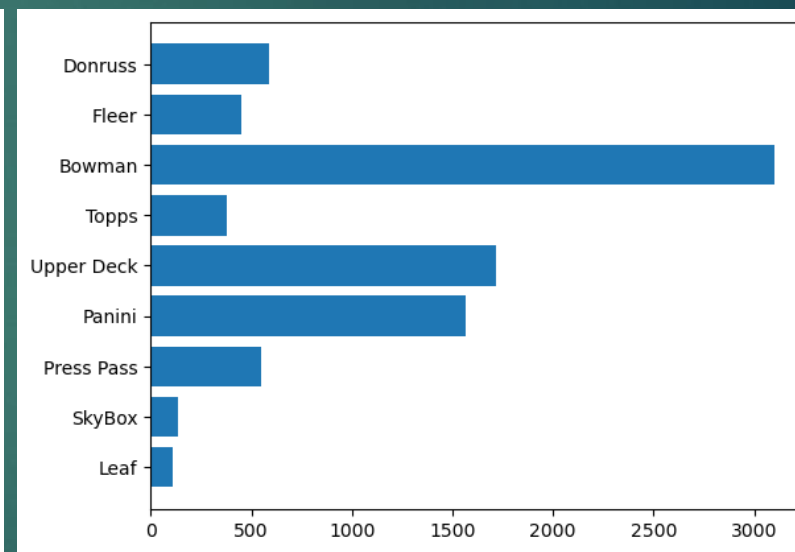
- אכן, ל-MT או MINT יש השפעה חיובית על ערך המחיר של קלף אבל לא בתחום בייסבול, באופן מפתיע, כאמור, אולי בגלל שבייסבול הוא ענף ספורט שהתחיל לפני שלושת ענפים אחרים ולכן יותר ישן ומכאן כמות קלפים ישנים של בייסבול יותר גדולה ומתרכזת בעיקר בתקופות של הונס או בייב רות' או לו גריג.

# ויזואליזציה

▶ ואחרון חביב, אם לחברות מסוימות שמייצרות את קלפים יש השפעה על מחיר של קלף? הבא נבדוק:



בסיסבול

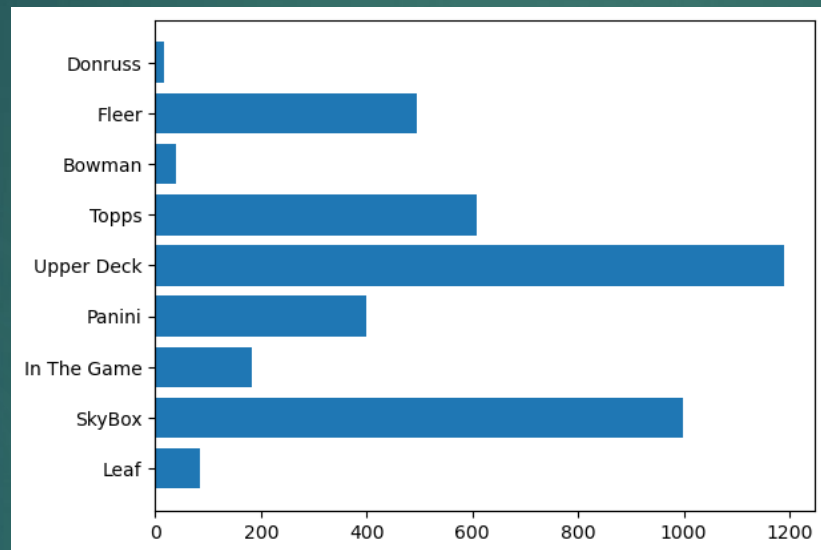


פוטבול

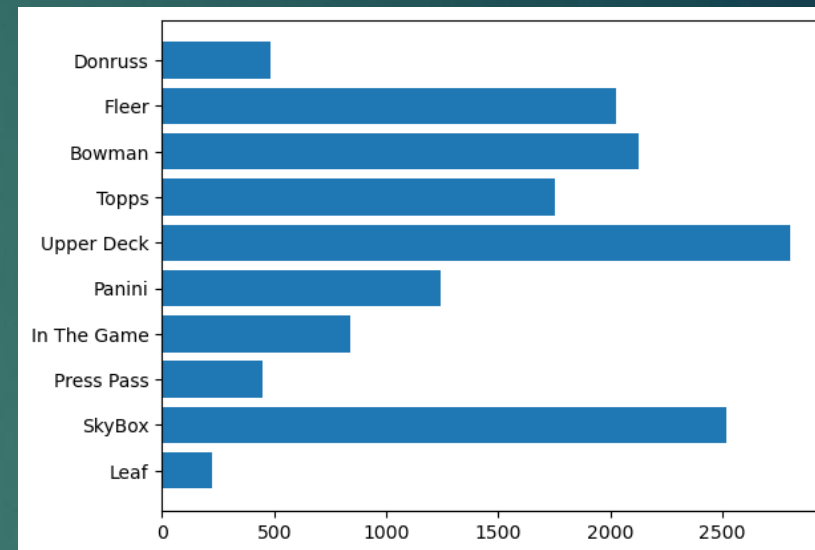
- הקדמה
- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- **ויזואליזציה**
- ניתוח נתונים
- מתקדם
- יישום והערכת
- ביצועים
- סיכום ומסקנות

# ויזואליזציה

- הקדמה
- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- **ויזואליזציה**
- ניתוח נתונים
- מתקדם
- יישום והערכת
- ביצועים
- סיכום ומסקנות



הוקי



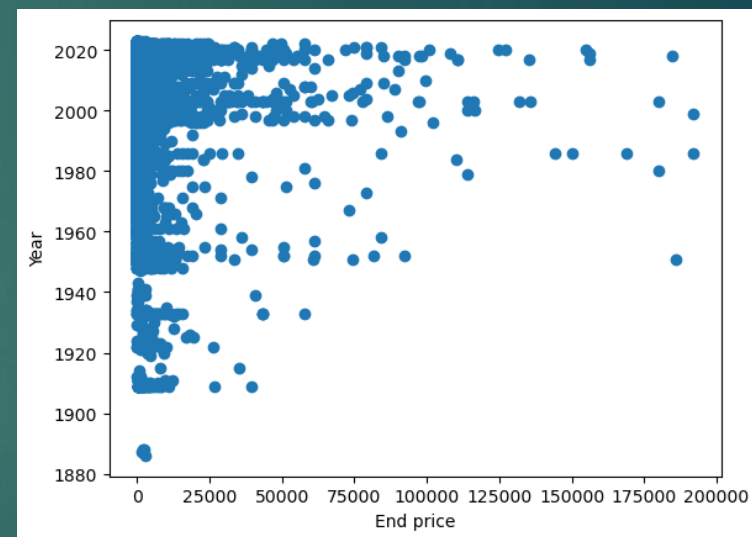
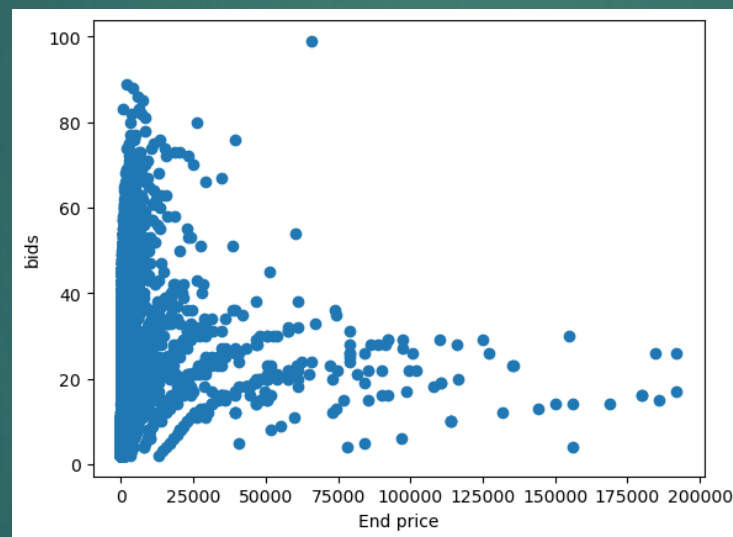
כדורסל

▶ ניתן לראות ש- Upper Deck ו- Fleer משחקים תפקיד משמעותי  
ו- SkyBox בחלק ענפי ספורט אבל החלטתי בכל זאת להשתמש ב- Panini  
בניתוח נתונים מתקדם כי הוא מאוד פופולרי בימינו



# ויזואליזציה

בנוסף, בדקתי אם קיים קשר בין מחיר מכירה סופית לשנת ייצור של קלף (אופן מציאת שנת ייצור קלף יתואר בשלב ניתוח נתונים מתקדם), כמו כן, אם קיים קשר מספר bids למחיר מכירה סופית ולהלן הגרפים:



כמו שאנחנו רואים בגרפים שהם Scatter plot, לא קיימים קשרים מובהקים בין מספר bids או שנת ייצור קלף לבין מחיר מכירה סופית. ממשיכים לניתוח נתונים מתקדם.

- הקדמה
- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- **ויזואליזציה**
- ניתוח נתונים
- מתקדם
- יישום והערכת
- ביצועים
- סיכום ומסקנות

# ניתוח נתונים מתקדם

- הקדמה
  - מקורות הנתונים
  - והרכשה
  - ניתוח ראשוני וטיוב
  - ויזואליזציה
  - **ניתוח נתונים מתקדם**
  - יישום והערכת ביצועים
  - סיכום ומסקנות
- ישנם נתונים שאינם קטגוריאליים כמו שנת ייצור או מספר bids ויש עמודה שהיא קטגוריאלית אבל לא בינארית:
- שנת ייצור של קלף, השתמשתי בRegular Expression כדי לשאוב שנת ייצור של קלף, הוא תמיד מופיע בכותרת, בתחילת כותרת וזה הקל עלי מאוד בעבודה של מציאת שנת ייצור קלף כך ייצרתי עמודה חדשה בשם year שמכיל ערכים בין 1880 (בערך) ל2023 (והיה נתון חריג שנובע מטעות הקלדה שהוא 22020 והסרתי את הרשומה).
  - מספר bids (הצבעות) – נשאב מדפי תוצאות של חיפוש באופן ישיר, נע בין 1 ל99 בערך.
  - כפי שציינתי לפני כן, לכל ענף ספורט יש מספר קטגורי: (1- פוטבול, 2- בייסבול, 3- כדורסל, 4 – הוקי) ולכן זו עמודה קטגוריאלית בעלת 4 ערכים אפשריים

# ניתוח נתונים מתקדם

- הקדמה
  - מקורות הנתונים
  - והרכשה
  - ניתוח ראשוני וטיוב
  - ויזואליזציה
  - **ניתוח נתונים**
  - **מתקדם**
  - יישום והערכת ביצועים
  - סיכום ומסקנות
- ▶ בנוסף, לאחר שביצעתי ויזואליזציה, ביררתי אילו מאפיינים בולטים ביותר שמשפיעים על מחיר קלף כמו צבע Platinum או חברת דירוג BGS וייצרתי עמודות בינאריות עבור כל ענפי ספורט ביחד שעונה על השאלה של כן או לא, אם קלף הוא Platinum או אם חברת דירוג BGS דירגה את הקלף, ועוד... כך יש לי עמודות בינאריות בנוסף לעמודות מספריות ועמודה קטגוריאלית בעלת 4 ערכים, הן משמשות לי בבניית מודל של למידת מכונה.
- ▶ בשאלת המחקר שלי – מה עושה את הקלף לקלף יקר? השאלה די מעורפלת בגלל שלא הגדרתי היטב את השאלה מה זה יקר ומה לא? החלטתי שזה רעיון טוב לחלק את מחירים לשתי קבוצות ביחס של 20:80 כך שקבוצה של 20 מכילה קלפים יקרים וקבוצה של 80 היא קבוצה של קלפים לא יקרים.

# ניתוח נתונים מתקדם

- הקדמה
- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- ויזואליזציה
- ניתוח נתונים
- מתקדם
- יישום והערכת
- ביצועים
- סיכום ומסקנות

► כך מצאתי את הגבול המחיר שמפריד בין שתי קבוצות אלו:

```
1 limit = 330
2 less_from_limit = "{:2.2%}".format(full_data_cln[full_data_cln['price_c
3 more_from_limit = "{:2.2%}".format(full_data_cln[full_data_cln['price_c
4 print('The limit is {}'.format(limit))
5 print('Cards which less from {}$: '.format(limit) ,less_from_limit)
6 print('Cards which more than {}$: '.format(limit) ,more_from_limit)
7
```

The limit is 330

Cards which less from 330\$: 80.00%

Cards which more than 330\$: 20.00%

► כלומר אם זה יותר מ-330 דולר, זה נחשב ליקר אחרת זה לא יקר

# ניתוח נתונים מתקדם

- הקדמה
  - מקורות הנתונים והרכשה
  - ניתוח ראשוני וטיוב
  - ויזואליזציה
  - **ניתוח נתונים מתקדם**
  - יישום והערכת ביצועים
  - סיכום ומסקנות
- ▶ משמעות הדבר – שיצרתי לעצמי עמודה נוספת שעונה על השאלה אם זה יקר או לא, הרצתי על כל הרשימה בבדיקה אם זה יותר מ330 או לא, 1- יקר, 0 לא יקר.
- ▶ עמודה הזו היא למעשה משתנה תלוי במחקר

# יישום והערכת ביצועים

- הקדמה
  - מקורות הנתונים
  - והרכשה
  - ניתוח ראשוני וטיוב
  - ויזואליזציה
  - ניתוח נתונים
  - מתקדם
  - **יישום והערכת ביצועים**
  - סיכום ומסקנות
- ▶ לאחר שיש לנו עמודה בינארית שהיא עונה על השאלה אם המוצר הוא יקר או לא, זאת אומרת שזו בעיית סיווג, ומכאן עלי להשתמש במודלים שמתאימים לבעיות סיווג.
- ▶ בגלל שרוב העמודות שמשמשות כמאפיינים ללמידת מכונה הן בינאריות בעיקר, אני מוצא בעצי החלטה כמודל מתאים ביותר, ובמידת הצורך, אשתמש גם בRandom Forest, בנוסף, בדקתי עם ובלי עמודות שאינן עמודות בינאריות, לדוגמה עמודה של שנת ייצור קלף שמכילה 109 ערכים ייחודים (בין שנים 1880 ל2023) ולעמודה bids יש 86 ערכים ייחודים, החשש היא שאם אשתמש בהם יהיה למודל overfitting אבל מסתבר שלא, הן תרמו לשיפור מודל.



# יישום והערכת ביצועים

▶ ולהלן הערכת ביצועים בשימוש מודל עצי החלטה ללא עמודות שאינן בינאריות:

```
Accuracy on training data= 0.8242333582647718
Accuracy on test data= 0.8303972079109191
```

▶ והחלטתי להרחיב את המאפיינים גם לעמודות שאינן בינאריות כמו שנת ייצור קלף, סוג ספורט ומספר bids ולהלן תוצאות:

```
Accuracy on training data= 0.9746530374802627
Accuracy on test data= 0.8770151238158551
```

▶ וואו איזה שיפור, ובוא נבדוק איך מדדים של accuracy, precision, recall ו-f1

- הקדמה
- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- ויזואליזציה
- ניתוח נתונים
- מתקדם
- **יישום והערכת ביצועים**
- סיכום ומסקנות

# יישום והערכת ביצועים

- הקדמה
- מקורות הנתונים
- והרכשה
- ניתוח ראשוני וטיוב
- ויזואליזציה
- ניתוח נתונים
- מתקדם
- **יישום והערכת ביצועים**
- סיכום ומסקנות

```
Accuracy score: 0.8770151238158551
Precision score: 0.6967943009795191
Recall score: 0.6620135363790186
F1 score: 0.6789587852494576
```

▶ התוצאות לא הכי טובות אלא טובות אבל אולי אפשר לשפר עם RandomForest, הבה נבדוק עם המודל:

```
0.9746530374802627
0.8882333388731927

5]: 1 print("Accuracy score: ", metrics.a
    2 print("Precision score: ", metrics.p
    3 print("Recall score: ", metrics.reca
    4 print("F1 score: ", metrics.f1_score

Accuracy score: 0.8882333388731927
Precision score: 0.7269487750556793
Recall score: 0.6903553299492385
F1 score: 0.7081796485137775
```

▶ יש שיפור קל עם המודל, ולכן הייתי הייתי משתמש במודל הזה כדי לבחון אם הקלף הוא יקר או לא

# סיכום ומסקנות

- **הקדמה**
  - **מקורות הנתונים**
  - **והרכשה**
  - **ניתוח ראשוני וטיוב**
  - **ויזואליזציה**
  - **ניתוח נתונים**
  - **מתקדם**
  - **יישום והערכת**
  - **ביצועים**
  - **סיכום ומסקנות**
- ▶ לאחר איחוד Datafreams והגדרת סוג ספורט כמספר קטגורי, ביצעתי ניתוח טקסט כדי למצוא מהן מילים נפוצות ביותר וחילקתי אותן באופן ידני לקטגוריות כך שאוכל למצוא מילים נוספות מהאינטרנט ששייכות לאותן קטגוריות.
- ▶ לאחר שבחנתי לאילו מילים מכל קטגוריה נחשבות כבעלות השפעה גדולה על מחיר מכירה סופית של קלף, תיעלתי אותן לצורך בניית מאפיינים ללמידת מכונה.
- ▶ השתמשתי באלגוריתמים של עצי החלטה וRandomForest כדי לבנות מודלים עבור נתונים, הגעתי לרמות דיוק טובות, אם לא הכי טובות אבל לדעתי מספקות כדי לענות על השאלה – האם הקלף הוא יקר או לא.