**מועמדות של סטודנטים לאוניברסיטה**

צבי ברגר: 206126575

ליאל שוקר: 208393744

הפרוייקט שלנו נמצא בגיטהאב - <https://github.com/BergerZvika/Data-Analysis-university-rating-prediction>

* **הגדרת התרגיל**

פרוייקט 5: חיזוי סיכויי הגשת מועמדות לאוניברסיטה

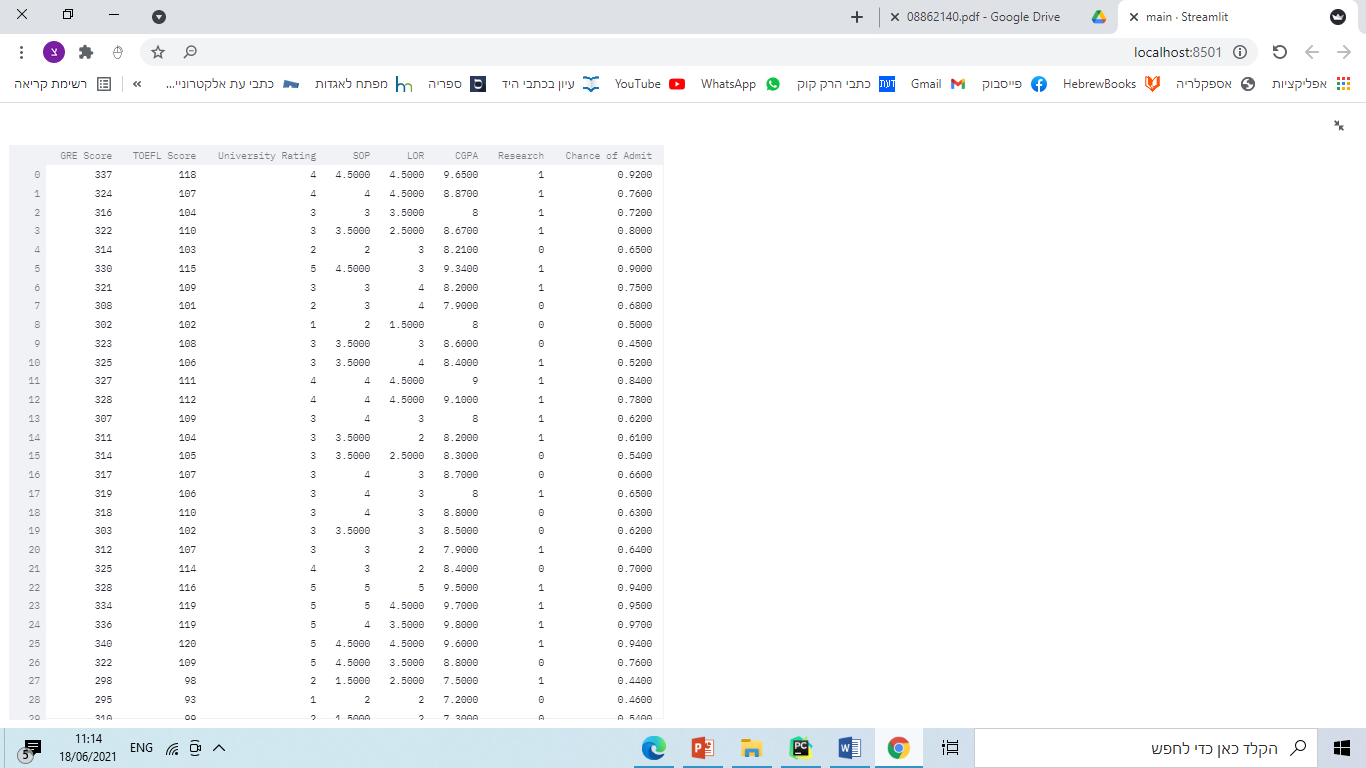
במאמר <https://ieeexplore.ieee.org/abstract/document/8862140> ניתחו החוקרים מה סיכויי הקבלה לכמה אוני' בהינתן נתונים אודות הסטודנטים. לצורך כך השתמשו בנתונים <https://www.kaggle.com/mohansacharya/datasets>. השאלה שלנו, בהינתן נתוני סטודנט, האם אפשר לבנות מודל ל"איזה אוניברסיטה הוא יגיש מועמדות"? האם מועמדים "פחות טובים" = "פחות מודעים למצבם ולכן יגישו לאוניברסיטאות לא ריאליות"? האם הם מתנהגים "נכון"?

* **מתודולוגיה**

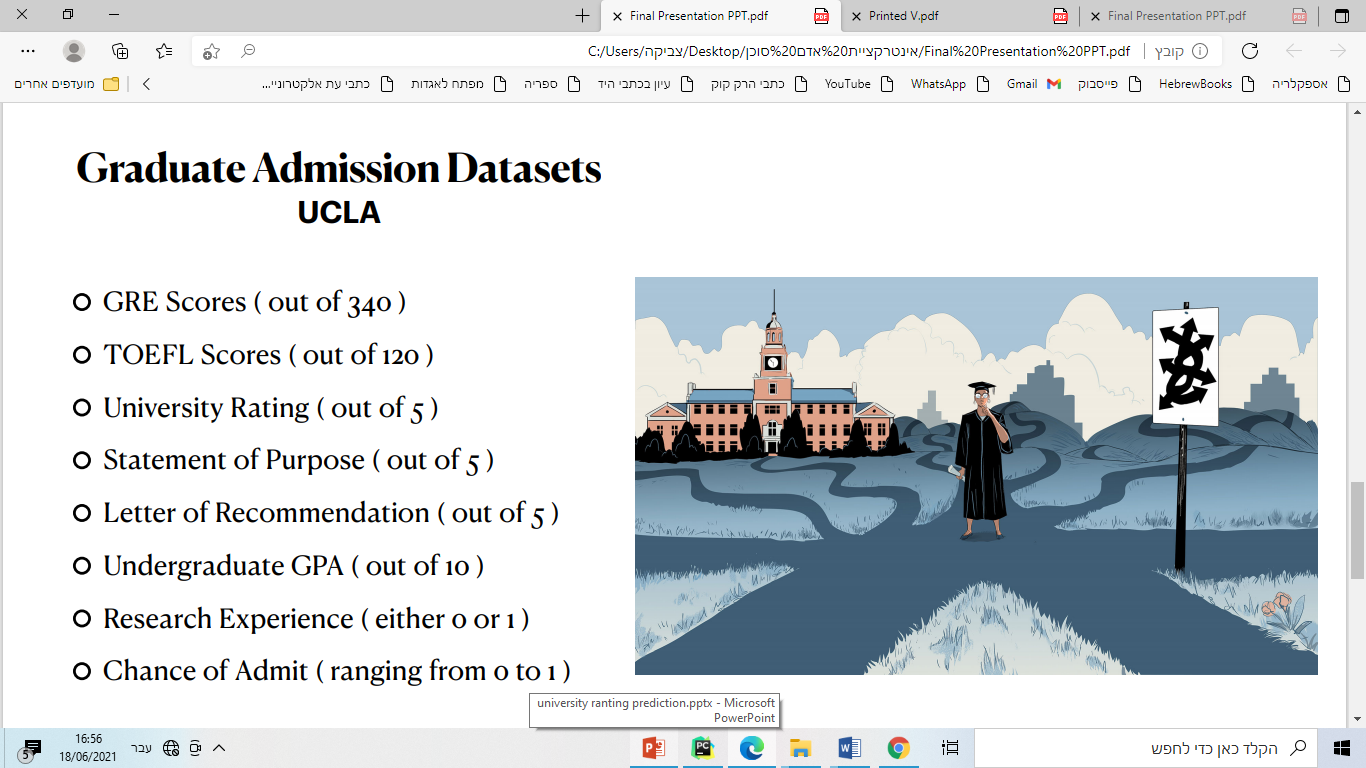
המודלים שבנינו לחיזוי הם מבוססים על ה-Data שקיבלנו (Data driven), ואין לנו שום כללים או עקרונות מבוססי מומחה (Expert Driven). ניסינו לחשוב האם ישנם באמת חוקים או כללים שאפשר להשתמש בהם על מנת להוסיף, אך מדובר פה בהתנהגות אנושית הנתונה להפתעות (לדוגמא גם אם למישהו יש את הציונים הכי טובים והוא עשה מחקר ויש לו המלצות אין זה ואמר בכלל שהוא רוצה לגשת ללמוד תואר שני גם אם סיכוייו גבוהים להתקבל, אז כל השאלה שלנו לאיזה רמה של אוניברסיטה הוא יגיש מועמדות תהיה שגיאה לחלוטין כי הוא בכלל לא רוצה להמשיך לתואר שני).

נשארנו עם ה-Data שלנו ואיתו ניסינו למדל את הבעיה, על כן המודלים שלנו הם מבוססי Data בלבד (Data driven).

* **קצת על הData-**

המידע שקיבלנו הוא ממחקר שנערך ומופיע בקישור kaggle לעיל. המידע עצמו נראה כך:

יש לנו 400 רשומות של סטודנטים.

כעת נסתכל עם הפיצ'רים של כל רשומה:

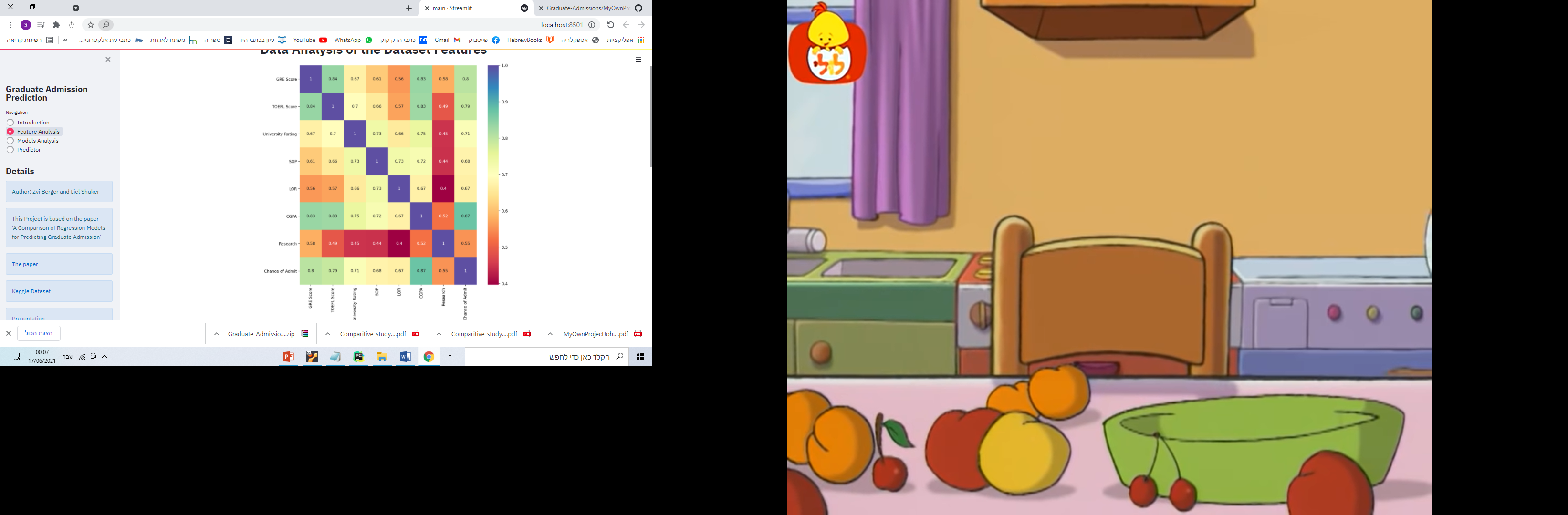
נחקור קצת את המידע שקיבלנו ונקבל את הטבלה הבאה:

נשים לב שכאשר נסתכל על הפיצ'ר CGPA, שמשמעותו הוא הציונים של האיש לפני לימודיו באוניברסיטה. מעניין לראות שהציון המינימלי אצלנו במידע הוא 6.8 והמקיסמום הוא 9.92, עם ממוצע של 8.5989. ניתן להסיק שאין לנו מידע על ערכים הקטנים מהמינימום שהוא 6.8 (יכול להיות שציונם לא מספיק כדי להתקבל ולכן לא הגישו מועמדות) ומכיוון שהממוצע הוא 8.6 שהוא מאוד גבוה זה מראה שמראש המידע שלנו מציג נתונים של אנשים מעל הממוצע. בעיה זו ניתן לראות גם בעוד מספר פיצ'רים אחרים כמו בפיצ'ר של GRE Score במידע שלנו המינימום הוא 290 והמקסימום הוא 340 עם ממוצע של 316.8076, ובפיצ'ר TOFEL Score שהוא עם מינימום 92 ומקסימום של 120 עם ממוצע של 107.41.

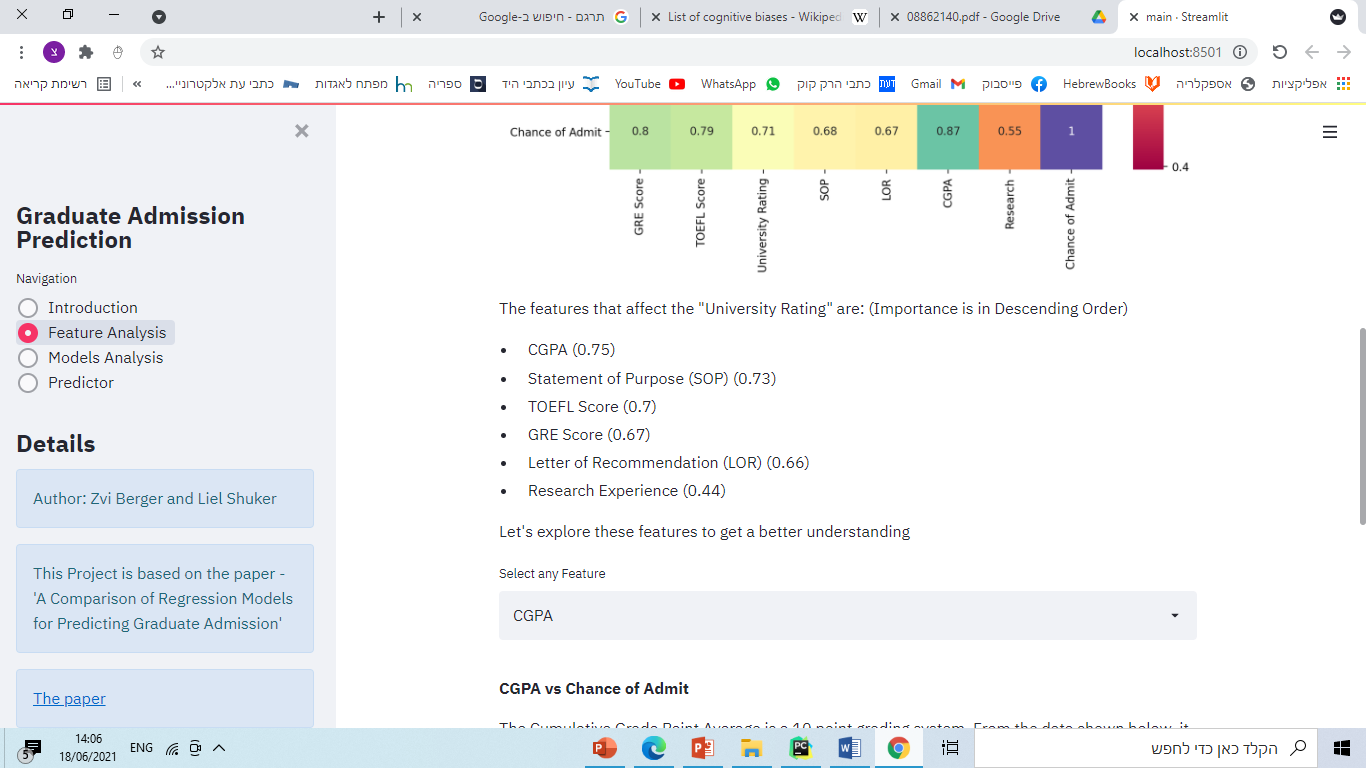
בנוסף המידע שלנו הוא מראש של אנשים שהגישו מועמדות לאוניברסיטה ובמאמר ניסו לבדוק איזה מודל חוזה בצורה המדוייקת ביותר את ההסתברות לקבלה. המידע שלנו לא מכיל אנשים עם ציונים גבוהים או נמוכים שלא הגישו בקשה לאוניברסיטה. ישנם אנשים שגם אם ציונים גבוהים מחליטים לא להגיש מועמדות מסיבותם שלהם והם אינם מיוצגים במידע.

לכן בגלל שמראש מדובר כאן במידע על אנשים שהגישו בקשה כלשהי לאוניברסיטה והם גם על הממוצע אנחנו נוטים לחשוב שיש כאם את בעיית selective labels problem, אנשים עם ציונים נמוכים פחות אינם מיוצגים אם בכלל והדבר נכון גם לאנשים שלא מגישים בכלל מועמדות.

* **חקירת הפיצ'רים**

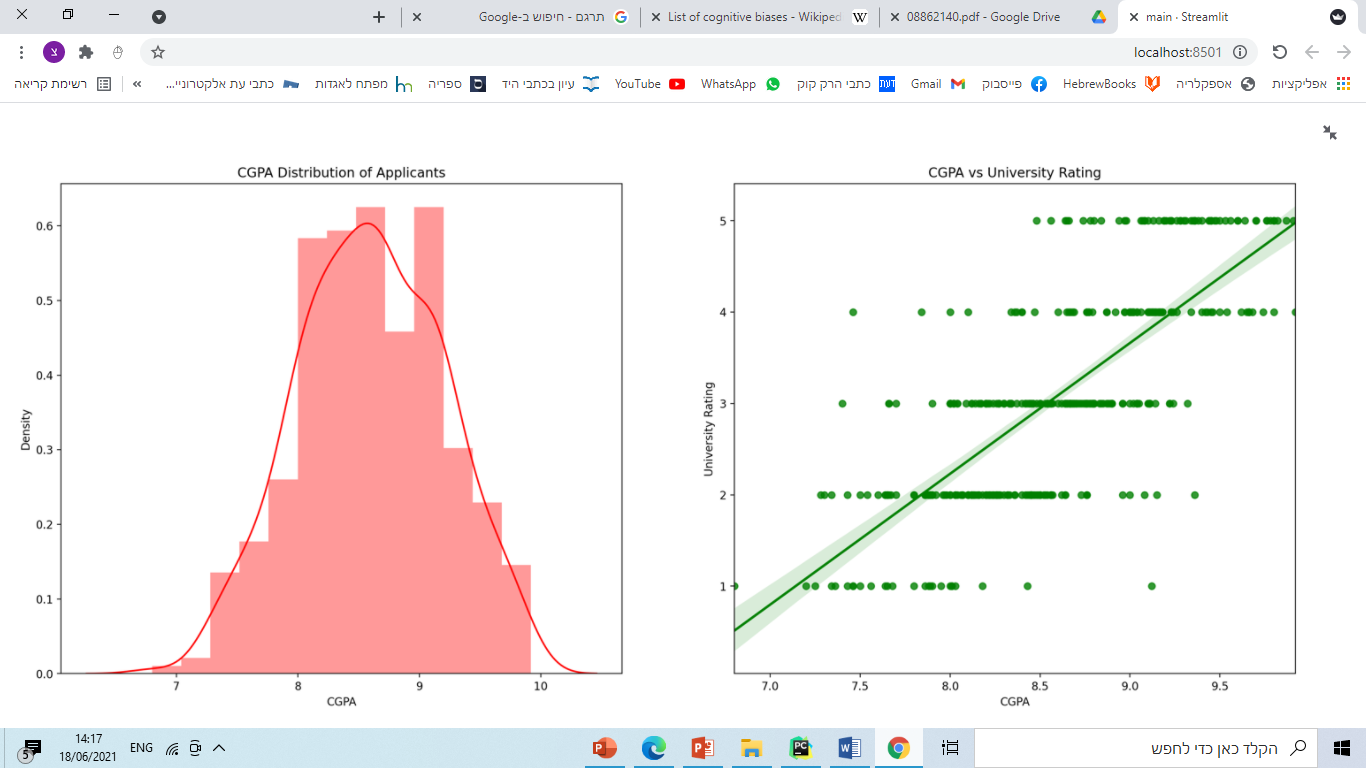
חקרנו את הפיצ'רים ויצרנו טבלת תלויות ביניהם. כל מקום בטבלה מראה את חוזק הקשר בין הפיצ'ר שבשורה לפיצ'ר שבעמודה. המספקים נעים בין 0 ל-1, כך ש-1 מראה על תלות מוחלטת וקשר חזק ו-ם מראה על אין שום קשר בכלל

נחפש בטבלה מהם הפיצ'רים שמשפיעים על university rating למען החיזוי. נקבל את הרשימה הבאה מהגדול קטן:



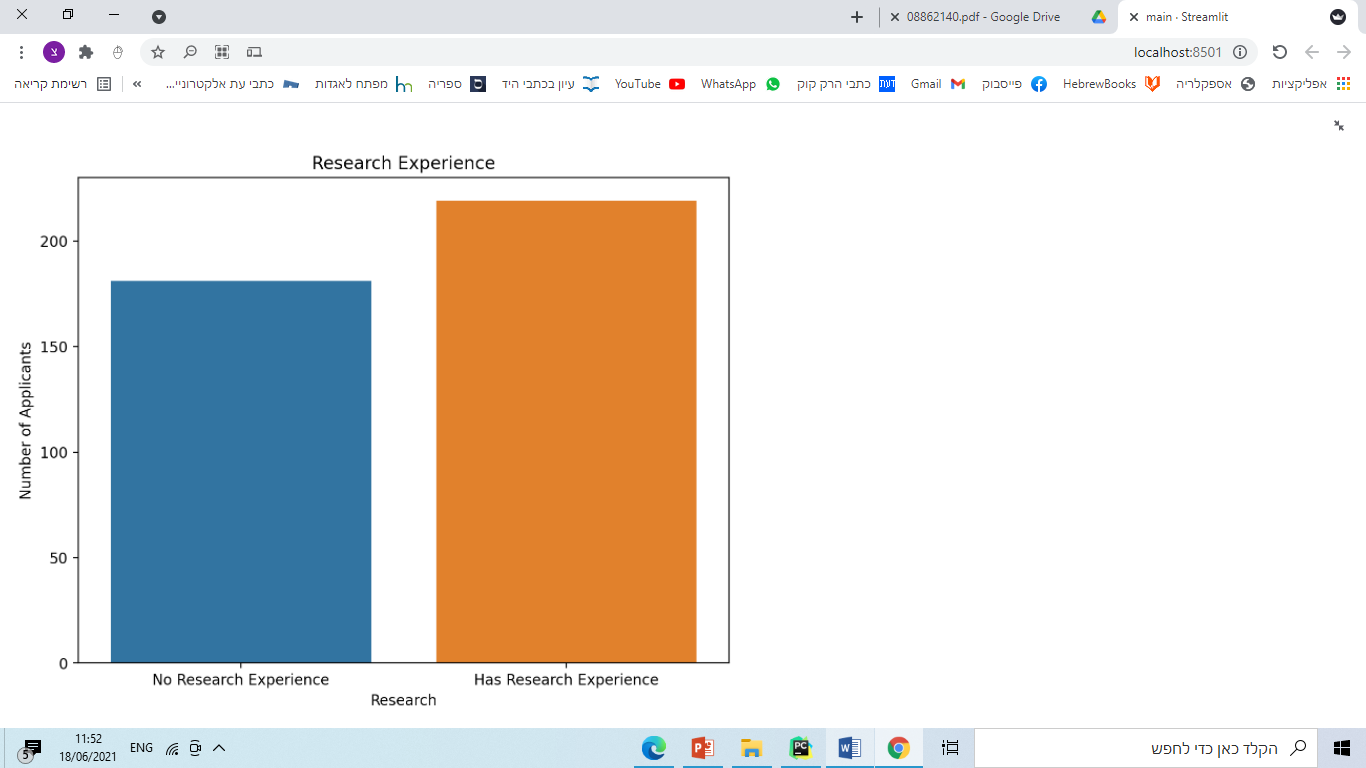
כלומר, הפיצ'ר שהכי משפיע על דרגת האוניברסיטה שאליה האדם יגיש מועמדות היא CGPA. בנוסף, ניתן לראות ששלושת המקומות הבאים בחוזק הקשר הם SOP, TOFEL Sscore, GRE Score שגם עליהם CGPA יש קשר חזק (0.72, 0.83,0.83 בהתאמה).

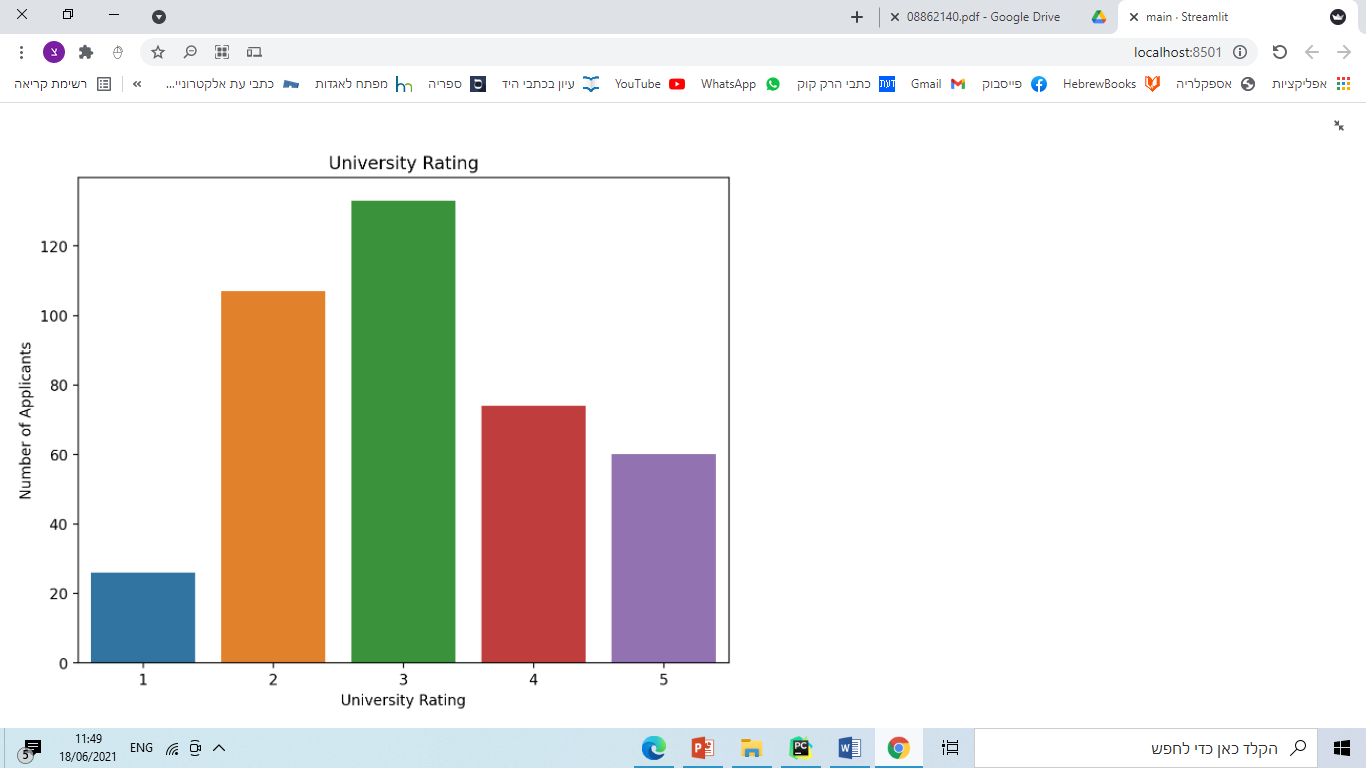
נסתכל על הפיצ'רים עצמם, כדי לקבל הבנה יותר טובה:

נתבונן קודם כל על הפיצ'ר CGPA, נראה שהוא מתפלג במידע כמעט נורמלי (גרף מצד שמאל) ורוב המוחלט של הציונים נע בין 8 ל9. ניתן גם לראות שיש קשר בין הCGPA לבין ה-university rating (גרף מצד ימין).

נסתכל כעת על הפיצ'ר GRE Score, נראה שגם הוא מתפלג פחות או יותר נורמלי (גרף מצד שמאל) וחוזק הקשר בין GRE Score ל-university rating חלש יותר (גרף מצד ימין):

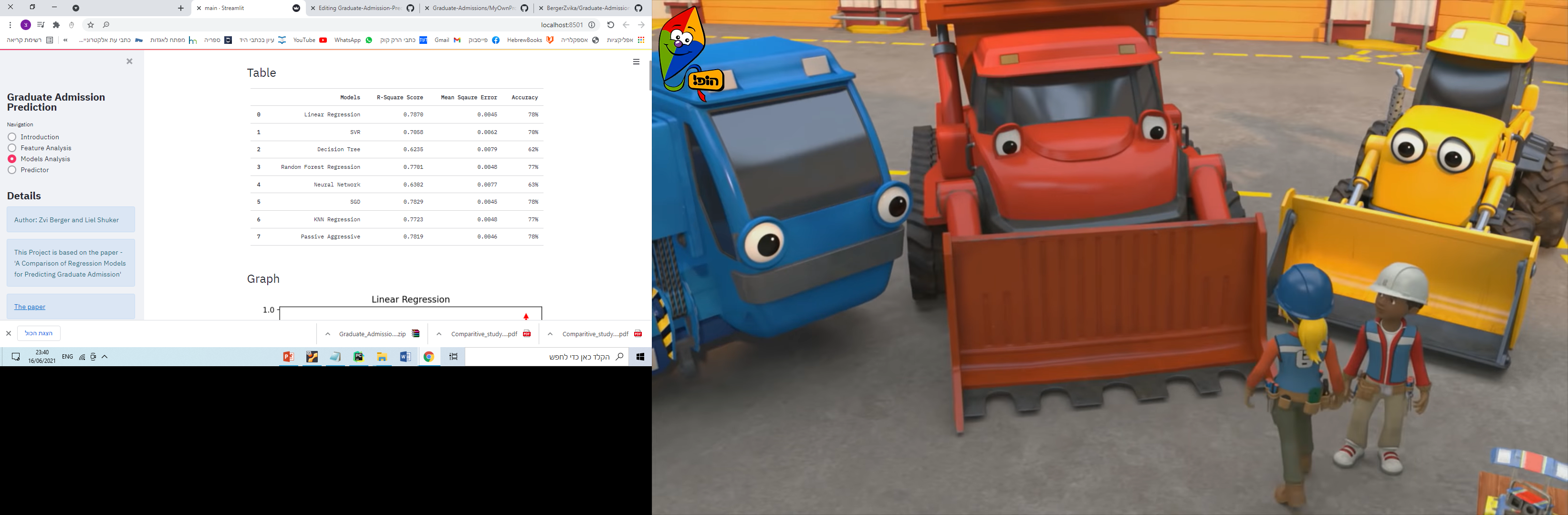
כעת, נסתכל על הפיצ'ר Research Experience. נשים לב שבמידע שלנו יש יותר אנשים שיש להם ניסיון במחקר מדעי לעומת אנשים שאין:



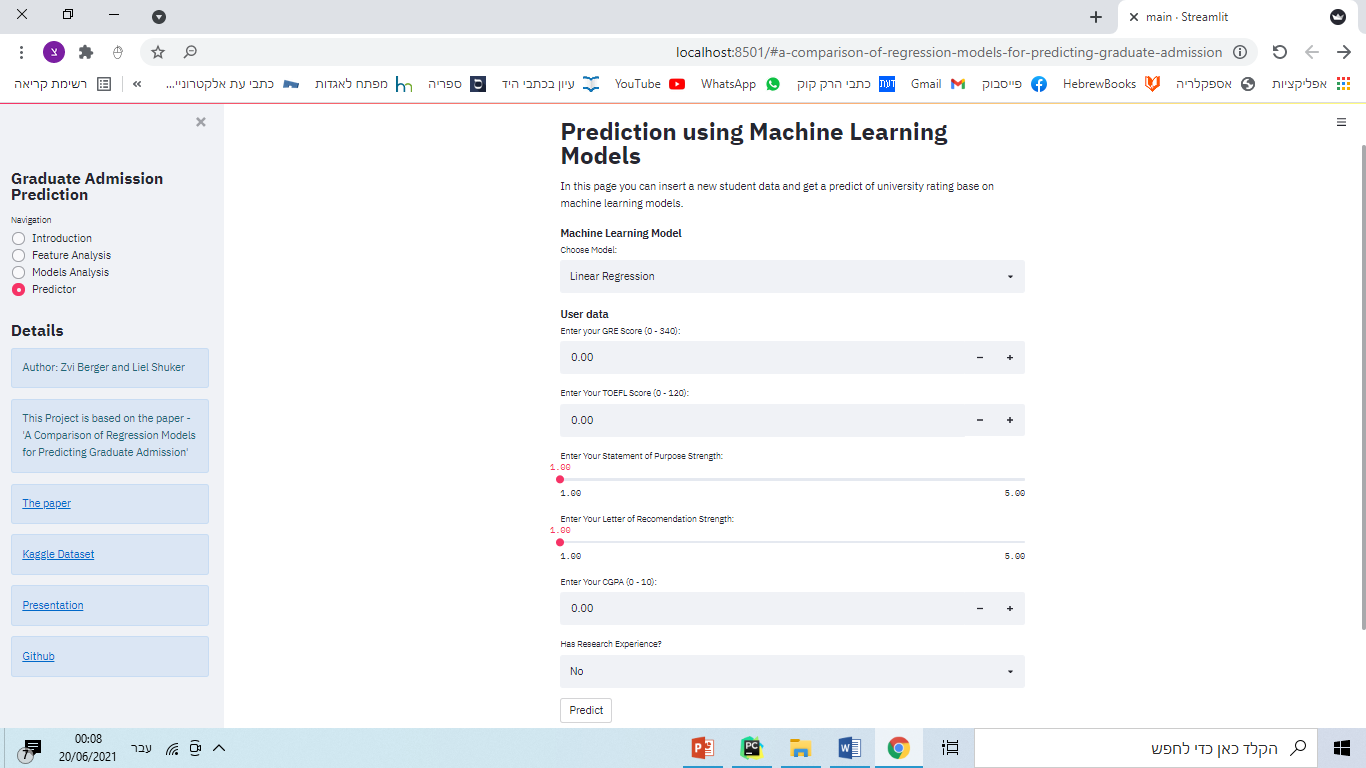
ולבסוף נסתכל על הפיצ'ר שאותו אנחנו חוקרים – university rating. ניתן לראות שבמידע שלנו רוה האנשים הציעו בקשות לאוניברסיטאות עם דירוג 2-3. בהינתן מה שכבר אמרנו שהמידע שלנו הוא גבוה מהממוצע היינו מצפים לראות שגם המועמדויות לאוניברסיטה יהיו בדירוג גבוה יותר (לפחות מהממוצע שהוא 3), לא רק שהוא לא הוא גם נוטה יותר לכיון 2. כנראה הדבר נובע מהנחה קוגנטיבית שלמדנו (cognetive bias) Loss Aversion. אנשים כנראה מעדיפים להגיש מועמדות לאניברסיטאות עם דירוג נמוך יותר ממה שהם יכולים כדי לא לחוש את תחושת ההפסד והכשילון בלא להתקבל.

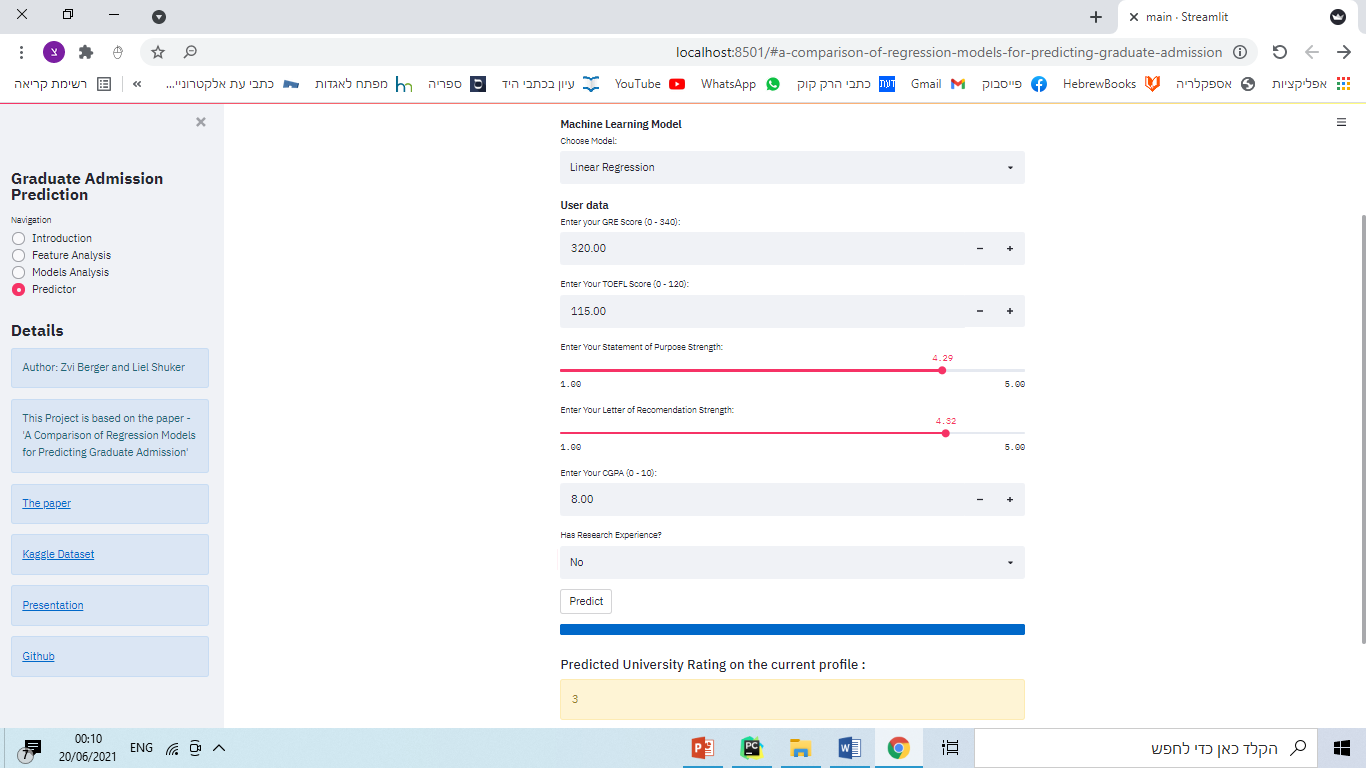
* **מודלים לחיזוי**

כדי לנסות לחזות מה האוניברסיטה שאליה מועמדות יגיש את מועמדותו השתמשנו במודלים הבאים:



חשוב לציין שרשת הנוירונים שהתשמשנו בה היא עם 3 שכבות חבויות בגדלים 30,50,30 עם פונקציית אקטיבציה relu, ו-shuffle.

לשם כך יצרנו עמוד להכנסת הפרטים ובחירת המודל:

אחרי הכנסת הפרטים ובחירת המודל ניתן לקבל חיזוי על ידי לחיצה על כפתור ה"predit" שלמטה:

ניתן לראות בתמונה לעיל שהמודל Linear Regression עם הפיצ'רים:

GRE Score: 320

TOFEL Score: 115

SOP: 4.29

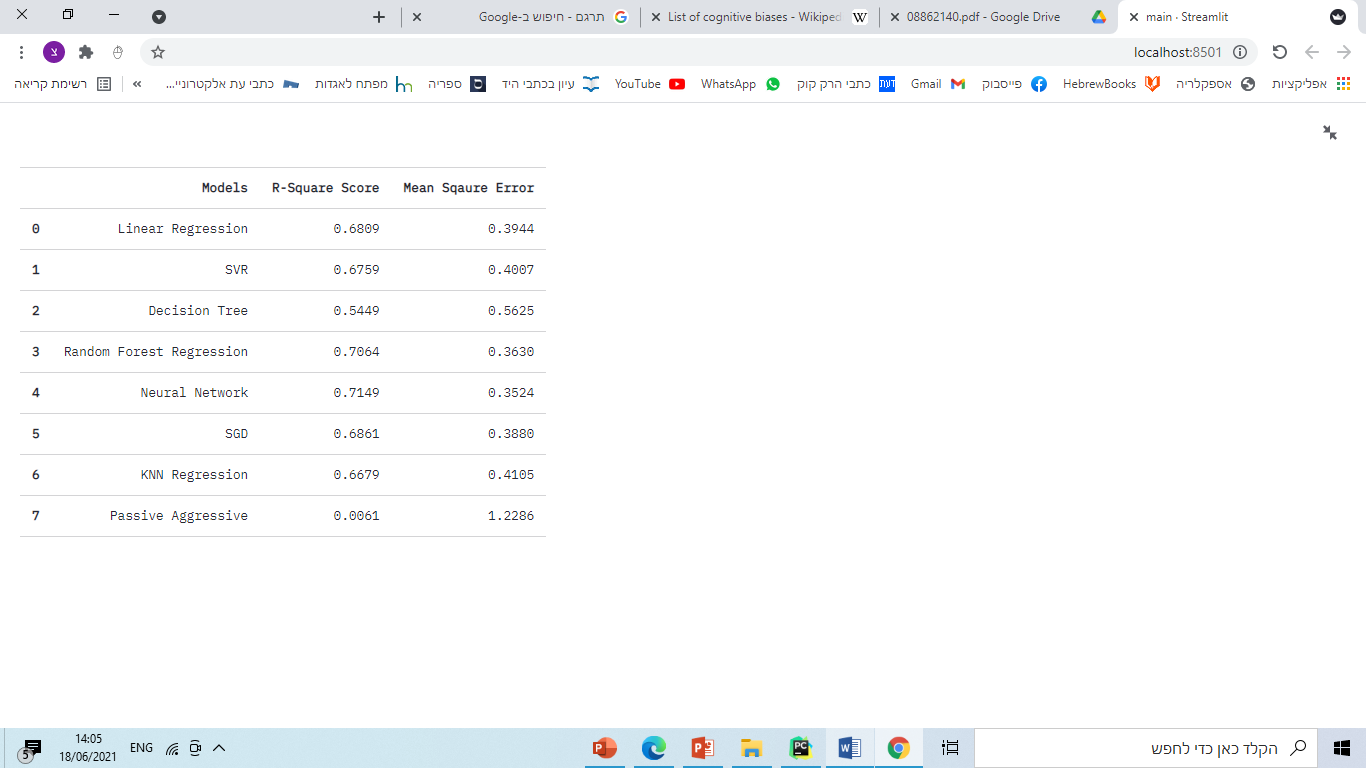
LOR: 4.32

CGPA: 8

Research Expirence: 0

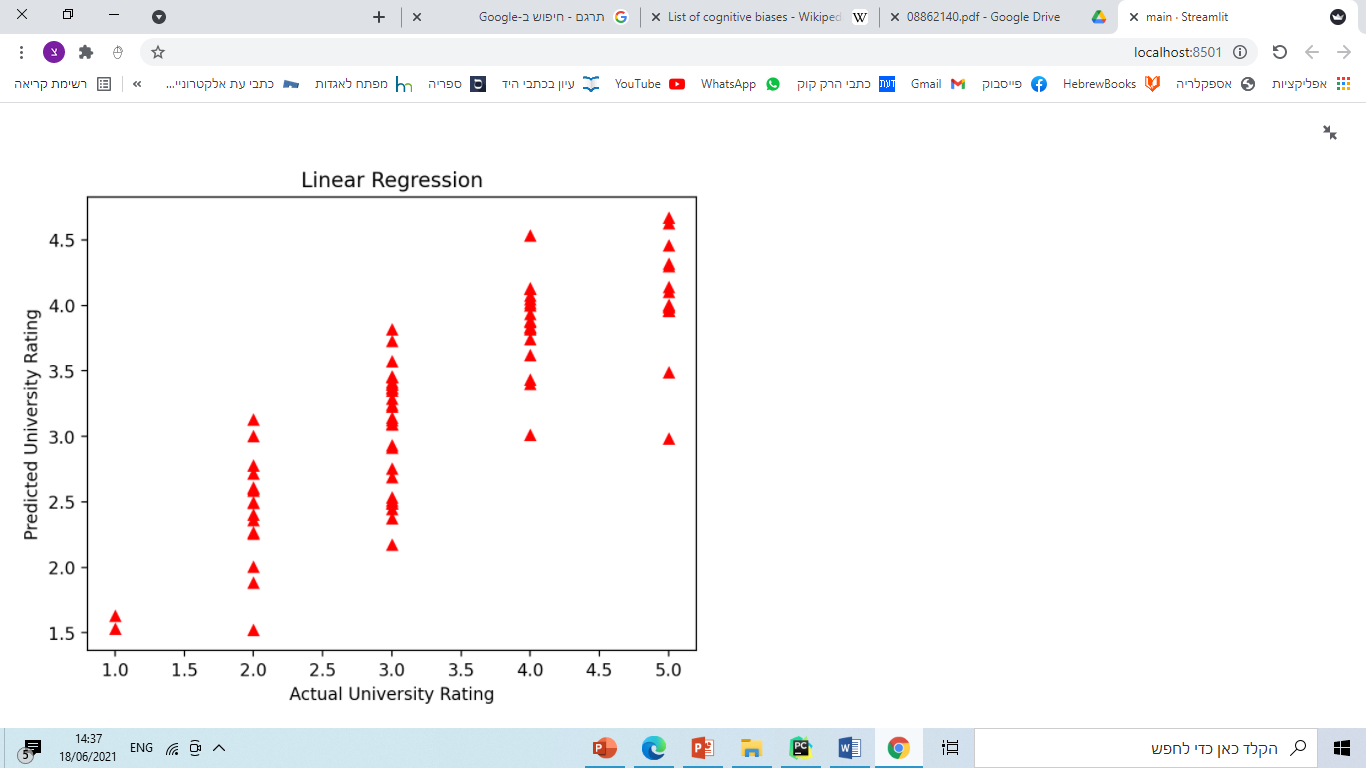
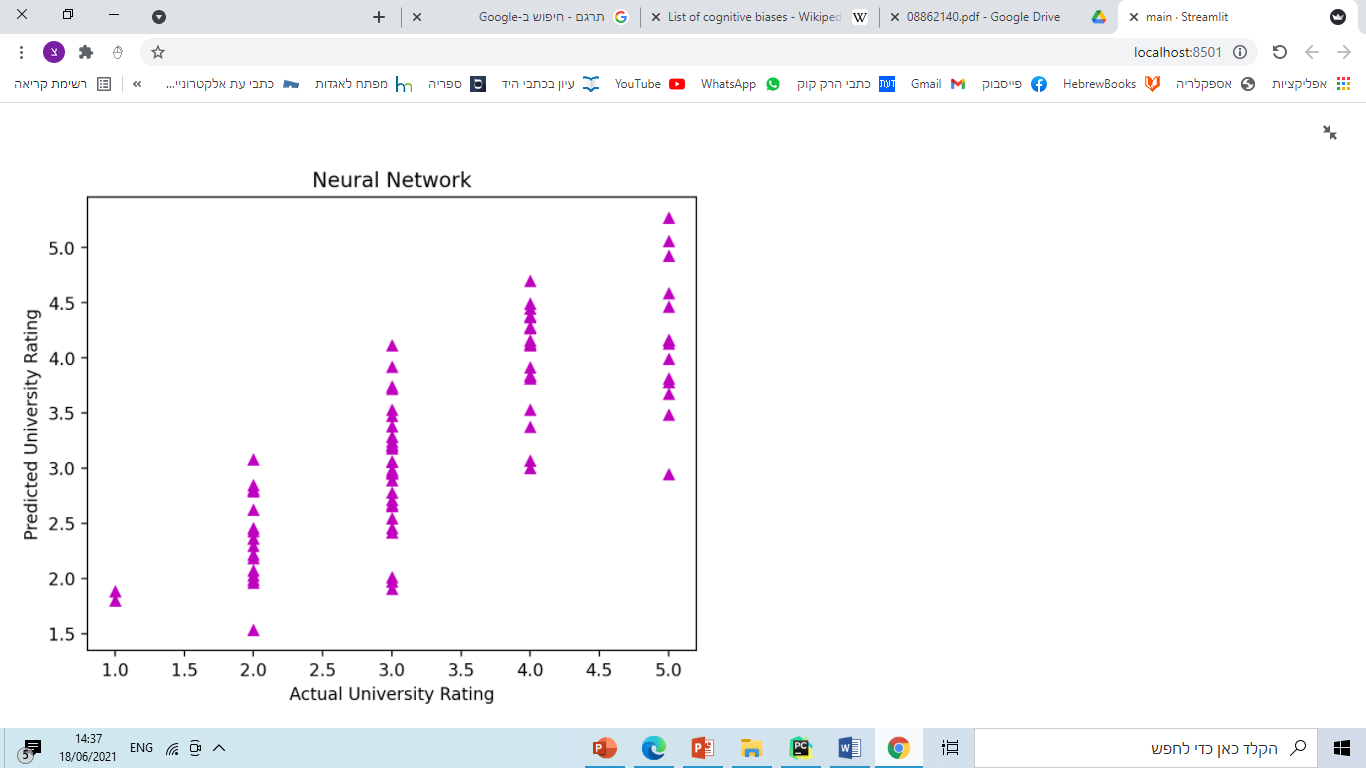
יקבל כחיזוי שהוא מגיש מועמדות לאוניברסיטה עם דרגה 3.

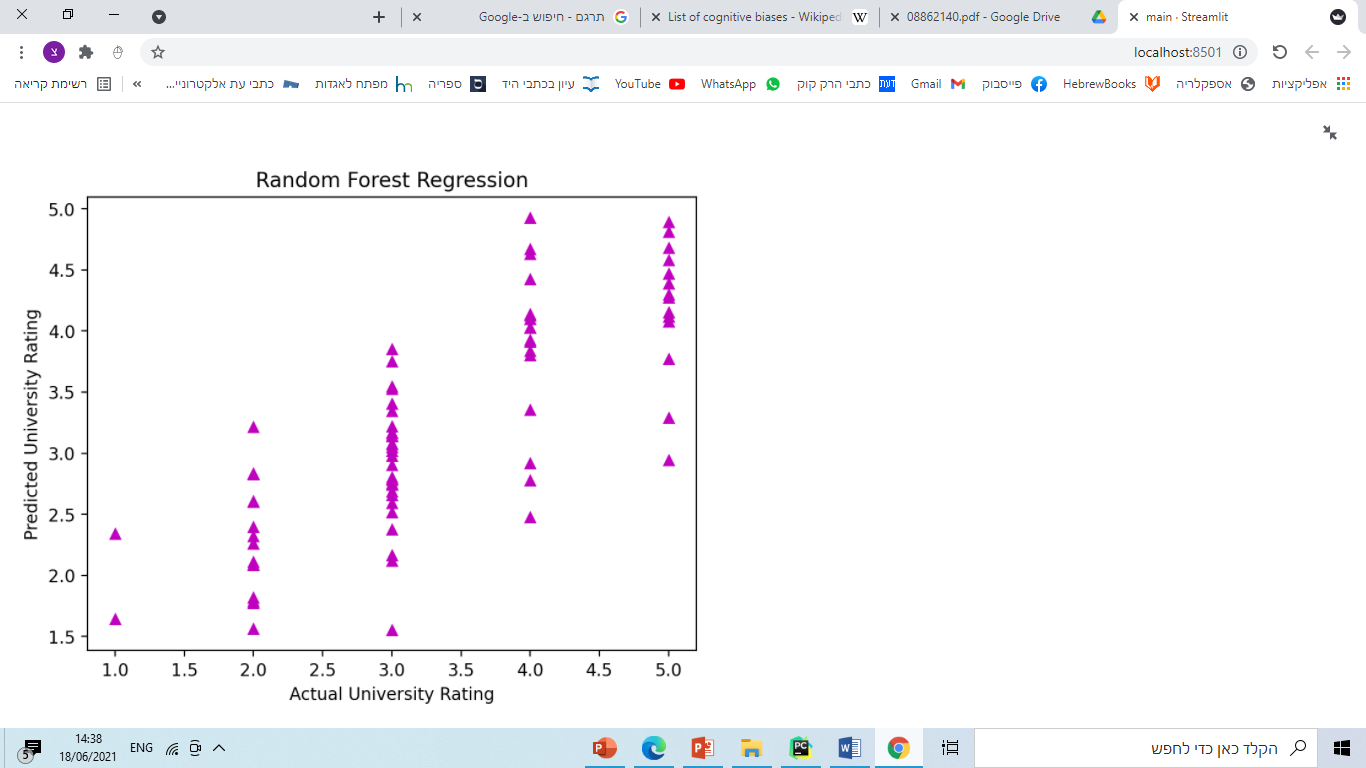
יצרנו השוואה בין המודלים על בסיס MSE ו-R2-Square וקיבלנו:



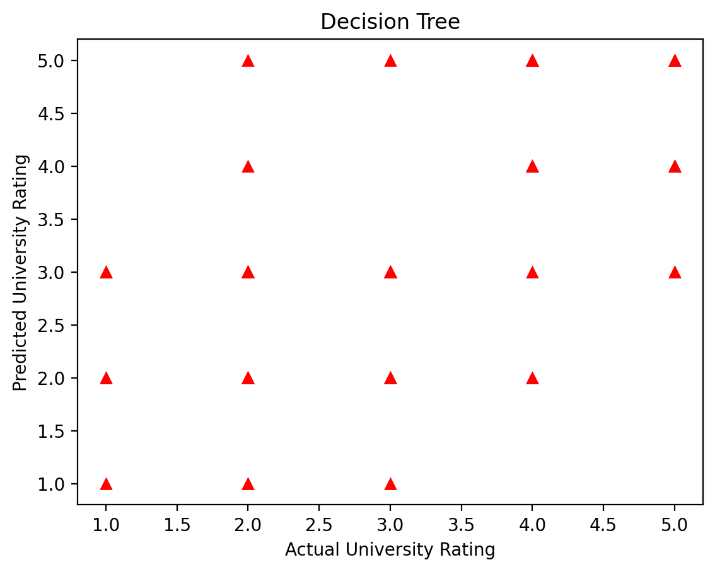
במאמר השוו בין המודלים: Linear Regression, SVR, Decision Tree, Random Forest וקיבלנו שנמודל Linear Regression הביא את התוצאות הטובות ביותר (הם בדקו חיזוי של סיכויי קבלה).

ניתן לראות לפי הטבלה במקרה שלנו, שהמודל Linear Regression אינו הכי טוב על אף שהוא אינו רחוק מאוד מהמודלים הטובים האחרים. בהרצה שלנו קיבלנו שהרשת נוירונים קיבלה את התוצאה הטובה ביותר אבל בגלל שהיא עובדת על shuffle כל הרצה יהיו תוצאות מעט שונות ולכן היא לא בהכרח תמיד תהיה הכי טובה. לכן המודל הבא אחריו שהוא הכי טוב הוא Random Forest שבמקרה שלנו עקף את ה Linear Regression.

נראה את גרפי החיזוי לעומת האמת למודלים הבאים:

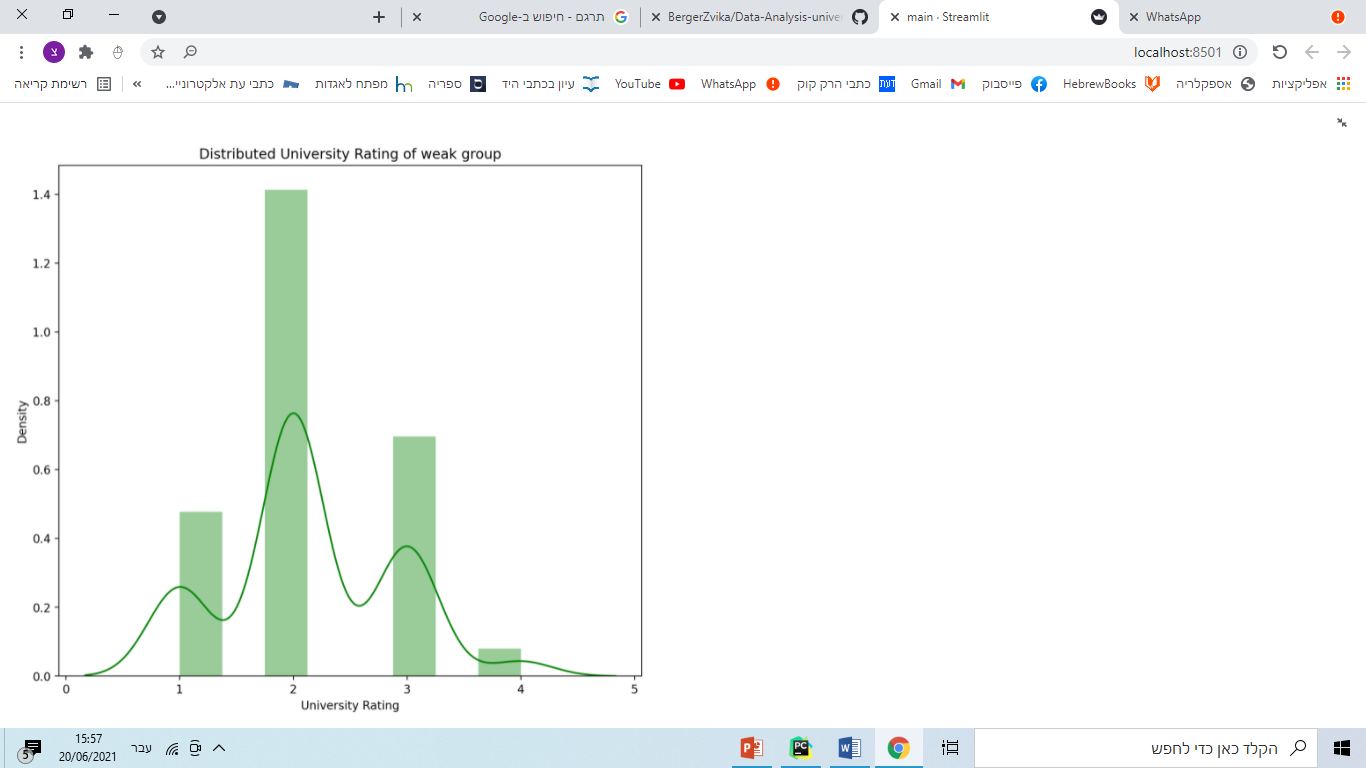


כמו כן אפשר לראות את הגרף של Decision Tree שהוא לא כל כך טוב כמו המודלים הקודמים שהראינו:

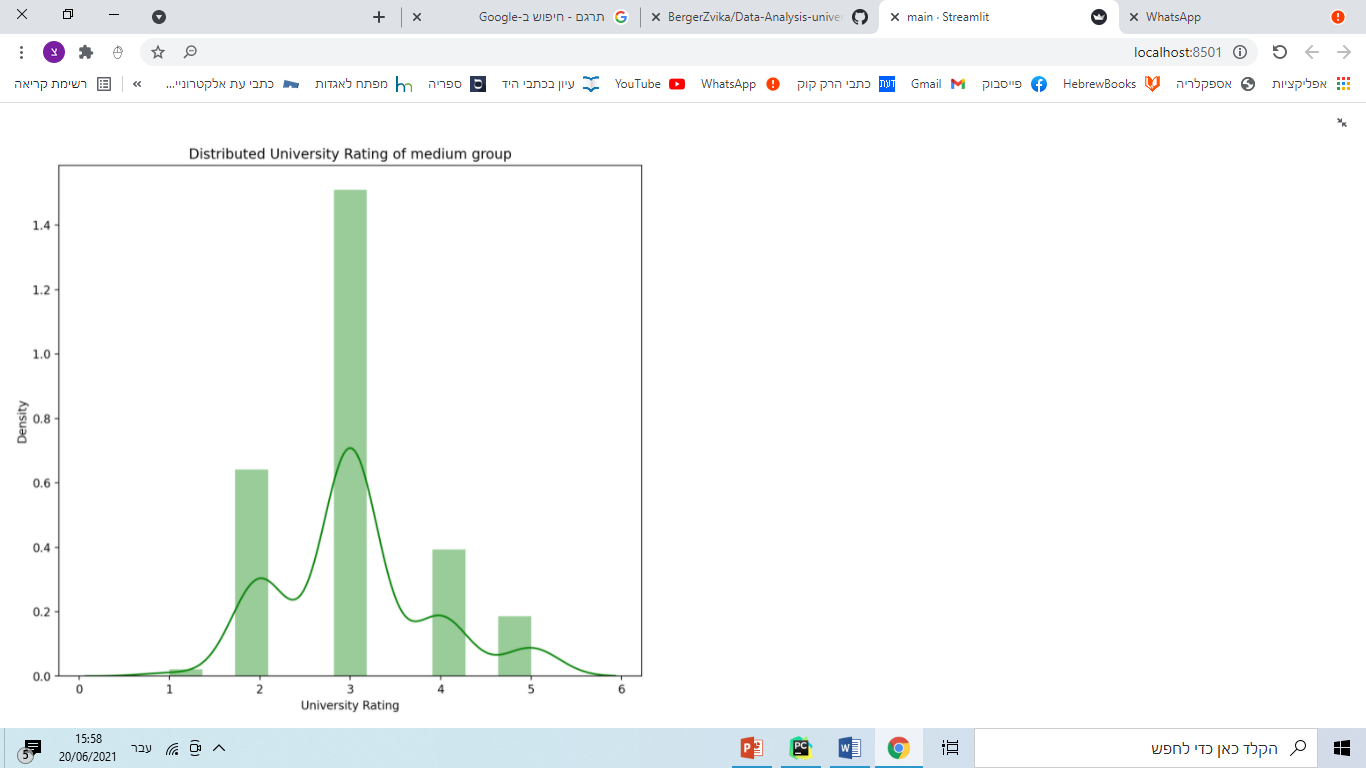


* **בדיקת רמות שונות של אנשים**

לצורך בדיקות בחלק זה חילקנו את המידע ל-3 קבוצות שוות. נציג עבור כל קבוצה את ההתפלגות של דירוג האוניברסיטאות שאליהם הגישו המועמדים בקשה.

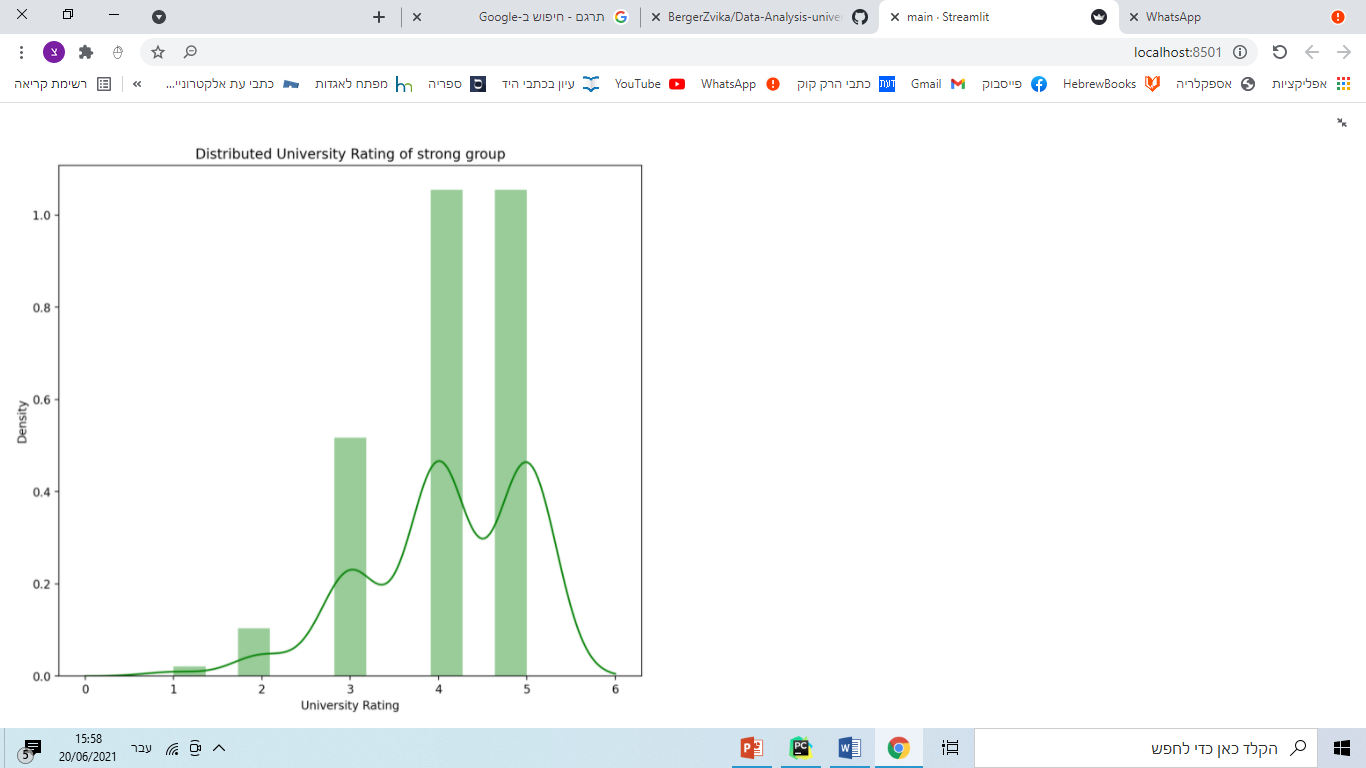
קבוצה ראשונה של השליש עם הציונים הנמוכים:

ניתן לראות שבקבוצה החלשה הגישו בעיקר לאוניברסיטאות בדירוג 2, לאחר מכן בדירוג 3 ו-1 וקצת אף הגישו לאוניברסיטאות בדירוג 4.

קבוצה שנייה של השליש עם הציונים באמצע:

ניתן לראות שהרוב המוחלט בקבוצה זאת הגישו מועמדות לאוניברסיטאות בדירוג 3. לאחר מכן ניתן לראות שהגישו יותר לאוניברסיטאות בדירוג 2 מאשר 4. לאחר מכן הגישו לדירוג 5 והכי פחות לדירוג 1.

קבוצה שלישי של השליש עם הציונים הגבוהים:



ניתן לראות שבקובצה החזקה הגישו הרבה לאוניברסיטאות בדירוג 4-5 כמצופה וככל שדירוג האוניברסיטה יורד כך גם כמות המועמדים.

נראה שסך הכל אנשים מגישים מועמדות כמצופה. החזקים מגישים לאוניברסיטאות חזקות הבינונים לבינוני והחלשים לאוניברסיטאות חלשות יותר. מה שמראה שאנשים מודעים לרמה שלהם ושל האוניברסיטאות השונות.

* **מסקנות**

לאור כל הבדיקות שעשינו הגענו למסקנות הבאות:

* + בדרך כלל אנשים בוחרים בצורה נכונה לאיזה דרגה של אוניברסיטה להגיש מועמדות.
  + ישנה נטייה להגשת מועמדות לאוניברסיטאות טיפה פחות טובות ממה שבאמת ניתן להגיש.
  + ישנו קשר בין רמת הציונים לבחירת דרגת האוניבריסטה.
  + כדאי להוסיף עוד data לקבלת תוצאות למגוון נרחב יותר של אנשים ולנסיון חיזוי של "לא יגיש מועמדות".