מגישה: חיה פנטליאט
מרצה: סימה סיגמן

פרויקט ניתוח נתונים וויזואליזציות

Python- Pandas

רקע ונתונים:

עמותת "בית המדרש" מחזיקה ברשת מוסדות לימוד לגברים מהמגזר החרדי בתחומי הלכה הנוגעים למצוות הארץ וקדושתה. תמורת שעות הלימוד מקבלים הלומדים מלגת קיום על פי קריטריונים שונים.

רשת המוסדות מפוזרת ברחבי הארץ וכוללת מספר סוגי מוסדות:

א. מוסדות לימוד יום- היקף לימוד של כ-4 שעות בשעות הבוקר/ הצהריים,
5 ימים בשבוע לאורך כל החודש.
מלגת הלימוד: 2,600 ₪ לחודש.
מספר המוסדות מסוג זה: 2 מוסדות.

ב. מוסדות לימוד ערב- היקף לימוד של כ-2 שעות בשעות הלילה,
5 ימים בשבוע לאורך כל החודש.
מלגת הלימוד: 800 ₪ לחודש.
מספר המוסדות מסוג זה: 10 מוסדות.

ג. מוסדות לימוד סוף שבוע- היקף לימוד של כ- 4 שעות.
פעם בשבוע, ביום שישי בבוקר, לאורך כל החודש.
מלגת הלימוד: בין 500 ₪ ל- 650 ₪, בהתאמה לשעות הלימוד.
מספר המוסדות מסוג זה: 3 מוסדות.

ד. מוסדות לימוד לפנות בוקר- היקף הלימוד של כ-2 שעות בשעות הבוקר המוקדמות,
5 ימים בשבוע לאורך כל החודש.
מלגת הלימוד: 600 ₪. תוספת של 250 ₪ למגיעים עד 5:50.
מספר המוסדות מסוג זה: מוסד אחד.
היות ומסוג זה יש מוסד אחד בלבד, אפשרויות הניתוח מועטות ולא יתבטאו בפרויקט זה.

ה. מוסדות לימוד חריגים- קיימים מוסדות נוספים בData base.
כללי הנוכחות ותשלום המלגה ייחודיים לכל מוסד.
מסיבה זו לא נציג את מוסדות אלו בניתוח הData.

הצורך:

עמותת "בית המדרש" מחלקת את מלגות הלומדים כיום על פי אחוזי נוכחות אך ללא נהלים ברורים לחלוקת המלגות, ובשל כך, מבקשת ליצור נהלי חלוקת מלגות אחידים לכל סוג מוסד.

בתור עובדת העמותה, התבקשתי ללמוד את הנתונים, לזהות את הכללים האחידים בכל סוג מוסד, לבודד את המקרים החריגים ולמקד אותם בכדי שתתאפשר עבודת ניסוח נהלים ברורים לחלוקת המלגות.

התהליך:

בשלב ראשון אספתי את נתוני תשלומי המלגות ואחוזי הנוכחות הקיימים במערכת ואיחדתי אותם לטבלת נתונים אחת. בכדי לאפשר הבנה טובה יותר של אופן חלוקת המלגות אספתי את נתוני הרבעון האחרון של שנת 2023.

מקורות המידע: מידע בנקאי (תשלומי מלגות), שעון נוכחות ודיווחי נוכחות.

חשוב לציין שסביב איסוף הנתונים הייתה הרבה עבודת פילטור ל-Data, אך עדיין ישנם נתונים חסרים או כאלו שאיכותם נתונה בספק. בשל כך, יובאו מסקנות על סמך נתונים חלקיים- יודגש כי איכותם מוטלת בספק ולאחר הסקת המסקנות תיבדק אמינותם בשטח.
מסיבה זו גם לא אבצע פעולות פילטור בפקודות הPython השונות.

בשלב שני העליתי את טבלת הנתונים לחוברת עבודה של Jupyter וחילקתי את הData לפי סיווג המוסדות למשתנים שונים.

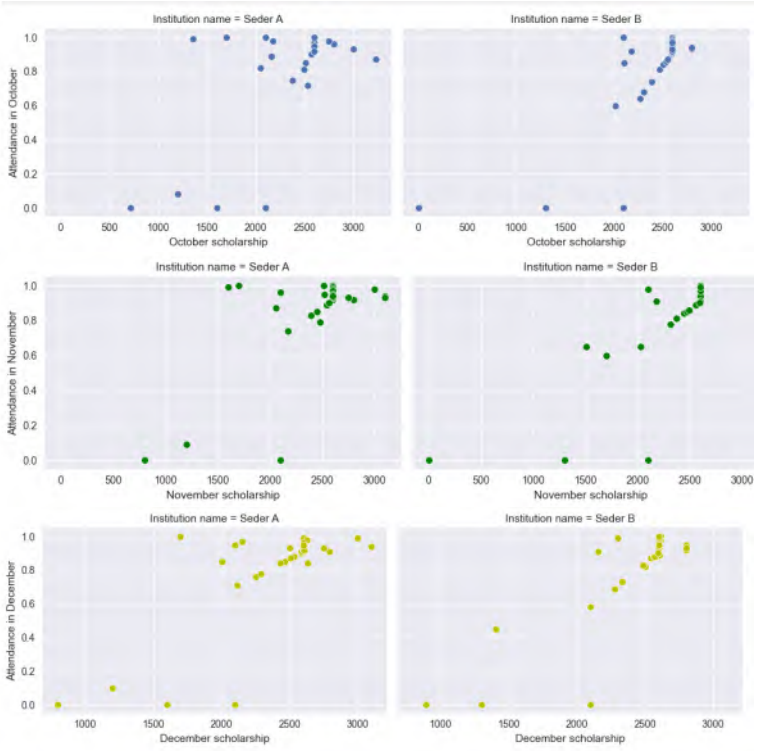
בשלב הבא, ביצעתי חישובים שונים על הData והמרתי את התוצאות לגרפים שונים לפי השלבים הבאים:

1. בדיקת אופן חלוקת המלגות והקשר בין נוכחות הלומדים לגובה המלגה.
את התפלגות הנתונים הצבתי בתרשים מסוג scatter, בחלוקה על פי המוסדות השונים.
2. בהתאמה למסקנות שעלו ומוצגות בפרק הבא- ניתוח נוסף ומפורט של הנתונים והתוצאות שעלו מתהליך החקר בכלים השונים.

מסקנות התהליך והסברן:

מוסדות יום (Type 1):

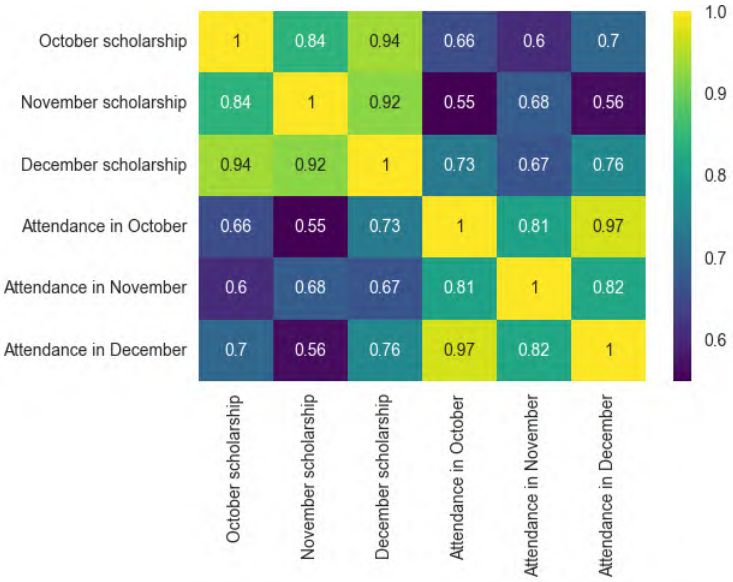
ניתוח הקשר בין נוכחות לגובה המלגה:



<u>תוצאות העולות מהגרפים:</u>

1. קיים קשר בין אחוז הנוכחות למלגה שמקבל הלומד.
2. מ90% נוכחות ומעלה הלומד מקבל מלגה מלאה.
3. במוסד "סדר א" הקשר בין אחוז הנוכחות לגובה המלגה פחות אחיד וחד משמעי.
4. בחודש דצמבר יש חריגות רבות יותר באופן יחסי לחודשים אחרים ברבעון.
5. בכל חודש ישנן מספר מלגות מועט המשולמות שלא על בסיס אחוזי הנוכחות של הלומד באותו החודש.

כדי לזהות את בעיית הקשר במוסד "סדר א" יצרתי קורלציה בין נתוני המלגות בחודשים השונים עם נתוני הנוכחות במוסד "סדר א" באמצעות Heatmap:

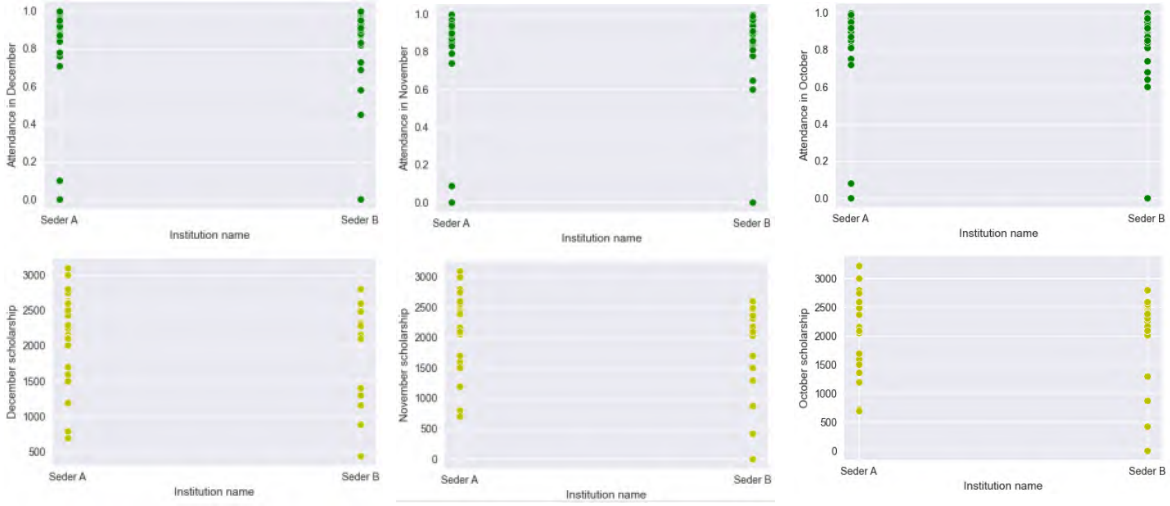


למעשה הקשר בין הנוכחות למלגה נמוך מאוד ועומד על שיעור של כ-70%.

בדיקה של אותם נתונים במוסד "סדר ב", בגרף זהה, מעלה קורלציה של למעלה מ80%.

פער זה מעיד שחלוקת המלגות במוסד "סדר א" נעשית במידה רבה ללא קשר לנוכחות.

כדי לבדוק את הנחה זו מכיוון נוסף, נבדוק את פריסת הנתונים מבחינת הנוכחות והמלגות בכל חודש, בחלוקה לפי מוסדות לימוד:



ניתן לראות כי התפלגות הנוכחות די זהה במוסדות השונים במהלך הרבעון, אבל, ללא קשר לפריסת הנוכחות, פריסת המלגות רחבה הרבה יותר.

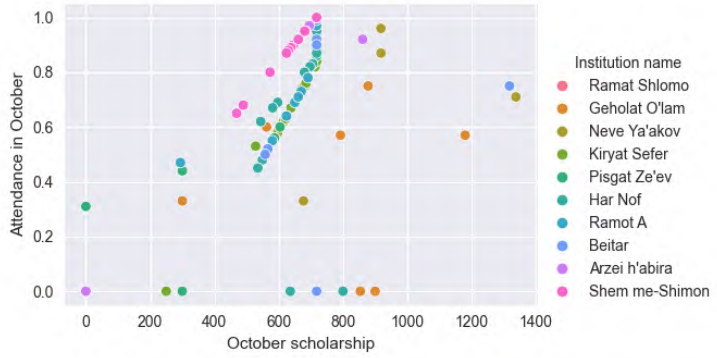
במוסד "סדר ב" פריסת המלגות זהה יותר לפריסת הנוכחות.

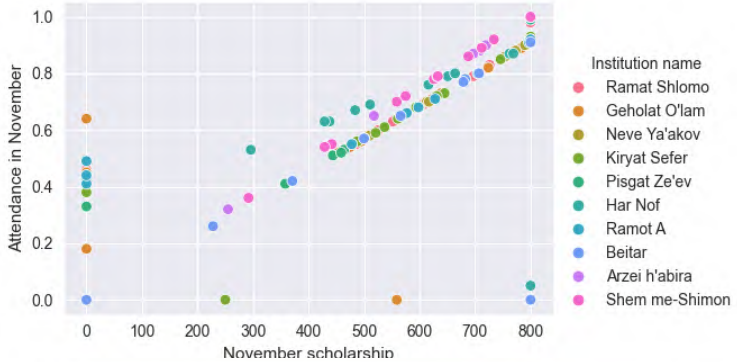
מסקנות:

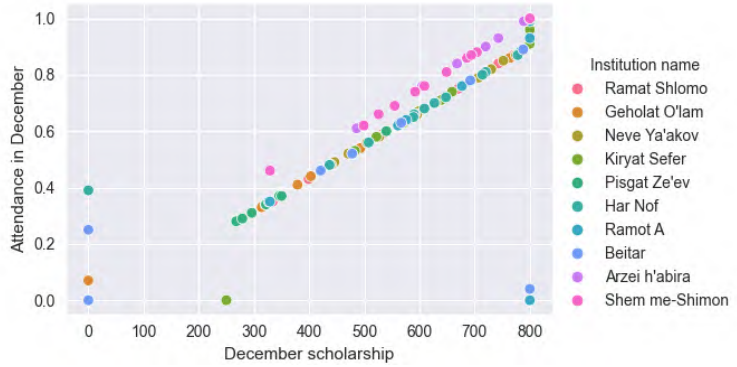
1. במוסד "סדר ב" קיים קשר ברוב נתוני הData בין גובה המלגה לאחוז הנוכחות. עם מעט חריגים.
2. במוסד "סדר א" אין כמעט קשר בין אחוז הנוכחות לגובה המלגה.
3. ניתוח הנתונים מעלה כי יצירת נהלים ברורים לחלוקת המלגה הינה הכרח מידי לצורך ניהול תקין.

מוסדות ערב (Type 2):

ניתוח הקשר בין נוכחות לגובה המלגה לפי החודשים השונים:







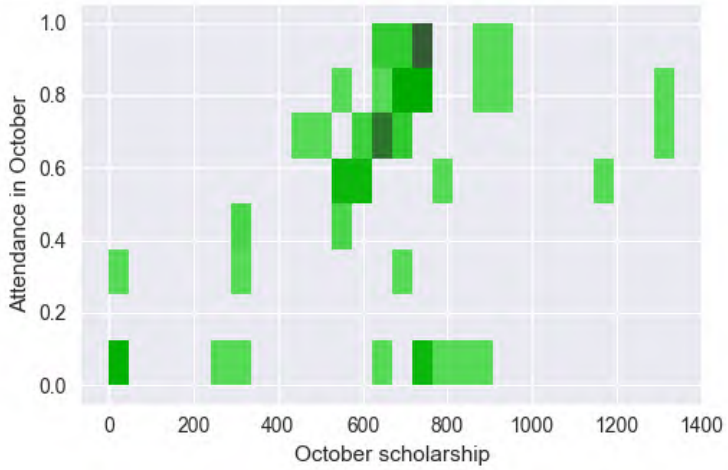
<u>תוצאות העולות מהגרפים:</u>

1. על פי רוב קיים יחס ישר בין אחוז הנוכחות לגובה המלגה. כאשר מ-90% נוכחות המלגה המתקבלת לרוב היא מלאה.

2. במוסדות "ארזי הבירה" ו"שם משמעון" המלגה הניתנת עבור הנוכחות גבוהה יותר ממוסדות אחרים.

3. בחודש אוקטובר אין קשר הגיוני כלל בין אחוז הנוכחות לגובה המלגה. גם קו המגמה שנוצר לכאורה- אינו הגיוני מבחינת היחסיות שבו.
בנקודה זו יש לסייג את מוסדות "ארזי הבירה" ו"שם משמעון" בהם מוצג קשר ישיר והגיוני כבחודשים העוקבים.

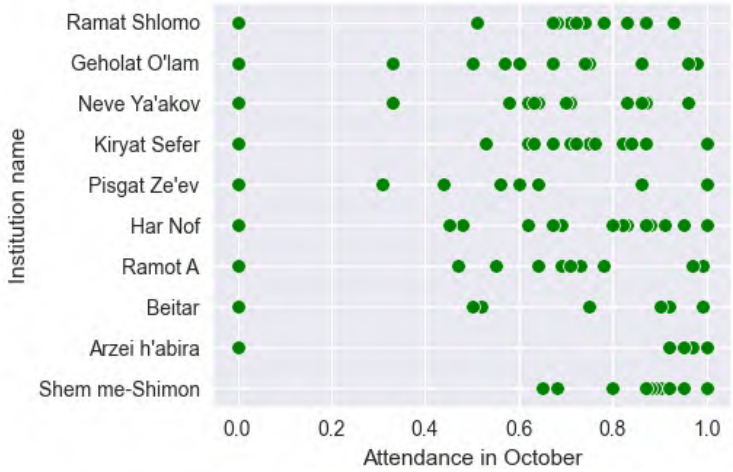
בחינת נתוני חודש אוקטובר:



בבדיקת היחס בין אחוז הנוכחות לגובה המלגה, מלבד העובדה שפיזור הנתונים אינו חד משמעי, מתחדדים נתונים המעלים שאלות:

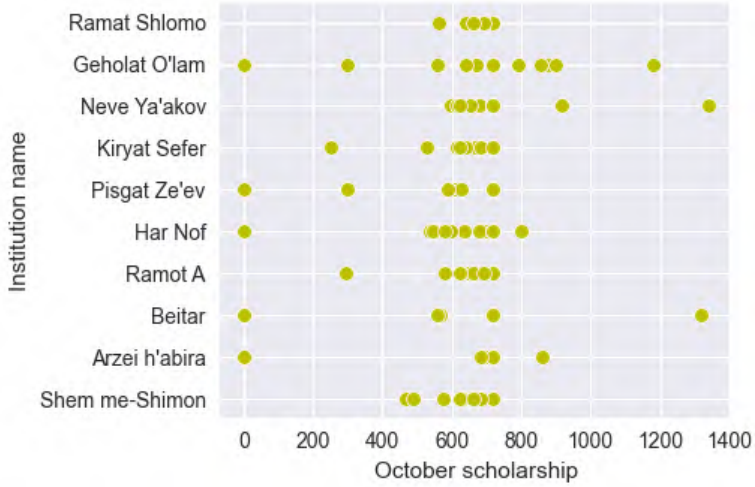
1. המלגות הגבוהות אינן באחוזי נוכחות גבוהים.
2. יש מלגות רבות שניתנו גם ללא נוכחות בפועל.

כדי להבין את ההתפלגות, בדקתי את השוואת הנתונים לחודש זה:



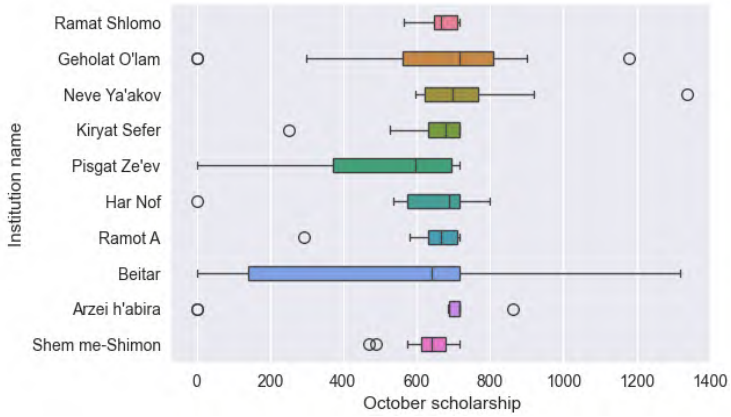
פיזור אחוזי הנוכחות בחודש זה נראה כהתפלגות נורמלית במוסדות השונים.

מרבית הלומדים נכחו ביותר מ-40% משעות הלימוד. נתון העולה גם מבחינת פריסת נתוני המלגות.



אך עדיין, למרות הצפי לראות את נתוני המלגות מתפלגים באותה רמת פיזור, הפריסה החזותית שונה מפריסת אחוזי הנוכחות.

גם מנתוני הסטטיסטיקה של סכומי המלגות בחודש זה ניתן לראות כי פריסת סכומי המלגות שונה בצורתה החזותית באופן ניכר מפריסת אחוזי הנוכחות: טווח הפיזור שונה בין המוסדות, סטיית התקן אינה אחידה ועוד.



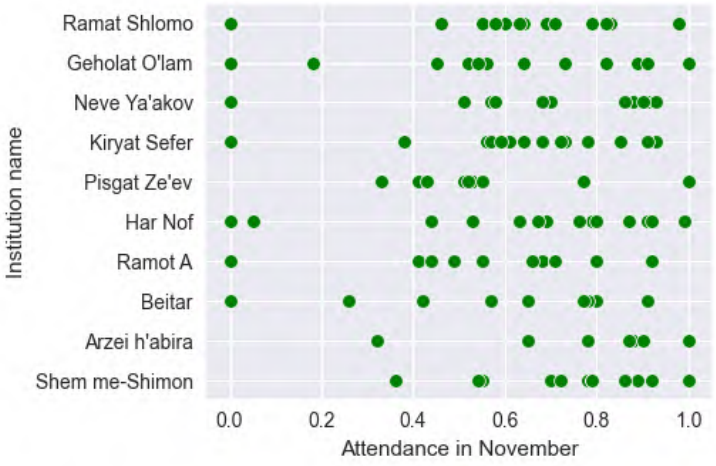
נתונים אלו מדגימים שוב עבורנו את הנתק הקיים בחודש זה בין אחוז הנוכחות לסכום המלגה.

קיימות 2 נקודות שיש לשים לב אליהן:
א. בחודש אוקטובר 2023 פרצה מלחמה.
ב. מרבית חודש אוקטובר הייתה חופשה עבור הלומדים בשל החגים.
יתכן כי נקודות אלו משפיעות על החלטת משלמי המלגות לתת משקל מופחת לאחוזי הנוכחות.

הנחה זו יכולה להסביר מדוע במוסדות "ארזי הבירה" ו"שם משמעון" קיים קשר דומה בחודש זה בין אחוזי הנוכחות לגובה המלגה כבחודשים הבאים: הקפדה על אחידות בנהלי חלוקת המלגות ללא שוני גם בזמני קיצון.

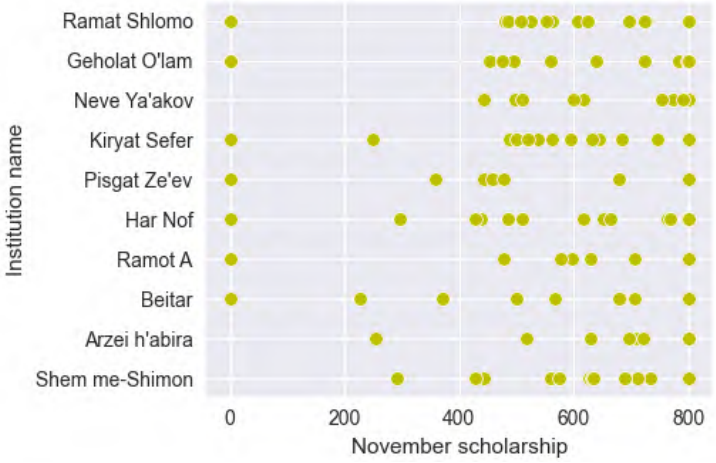
נתון זה מחייב יצירת נוהל אחיד בכל המוסדות מהו האופן בו מחלוקת המלגה בכל חודש, גם כאשר ישנם תרחישים לא צפויים המשנים את הנתונים בשטח או בחופשות ידועות מראש.

כדי לאשש את מסקנה זו, הצגתי גם את נתוני החודש העוקב- נובמבר, באותו גרף:



ניתן לראות כי בחודש זה פריסת אחוזי הנוכחות זהה ברובה לפריסת גובה המלגות במוסדות השונים.

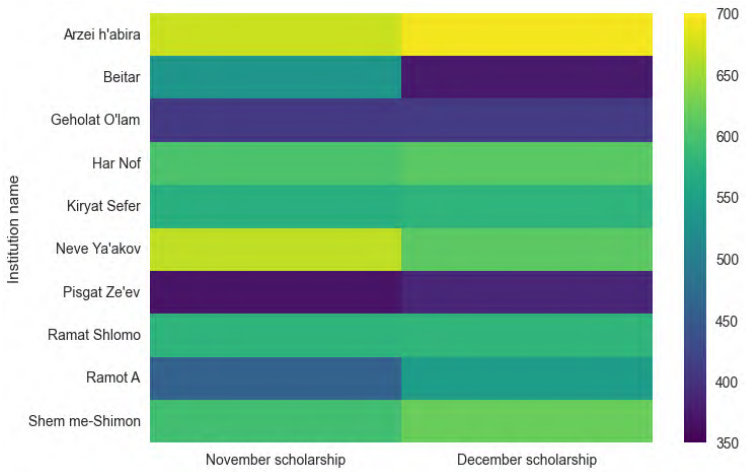
כך שבדיקת נתוני חודש זה מאשרת את המסקנה שהעליתי לגבי חודש אוקטובר.



לגבי המוסדות בהם חלוקת המלגות אינן באותו יחס כמו רוב המוסדות מסוג זה:

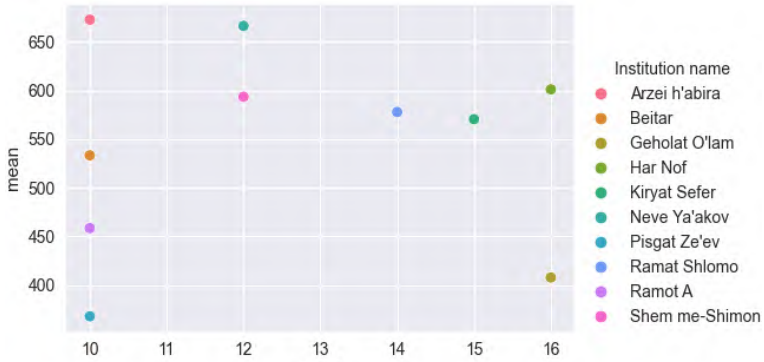
בשלב ראשון בדקתי את השוואת ממוצעי המלגות ללומד על פי מוסד בחודשים נובמבר-דצמבר:

בדיקה זו אינה יכולה להצביע על החריגים שהזכרנו בתחילת הניתוח (ארזי הבירה ושם משמעון), כי ממוצע המלגה אומנם גבוה, אך לא באופן חריג ביחס למוסדות אחרים בהם יחס הנוכחות- מלגה תואם לקו המגמה העיקרי שנוצר.

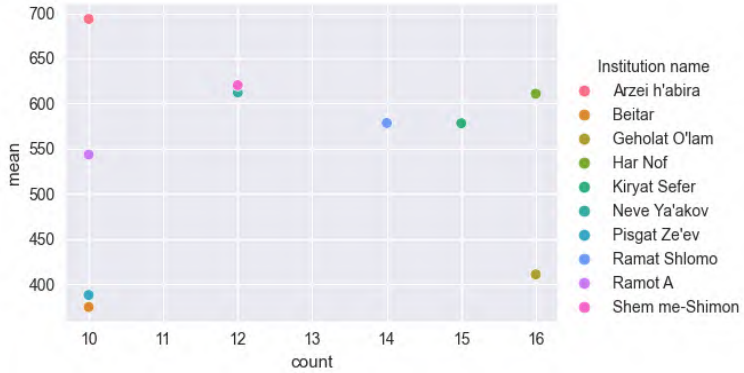


בשל כך, בדקתי את ממוצע המלגה ביחס לכמות הלומדים בכל מוסד בחודשים נובמבר ודצמבר:

גרפים אלו מציגים כי מרבית נתוני המוסדות אחידים בחודשים השונים מבחינת ממוצע המלגות ביחס לכמות הלומדים.

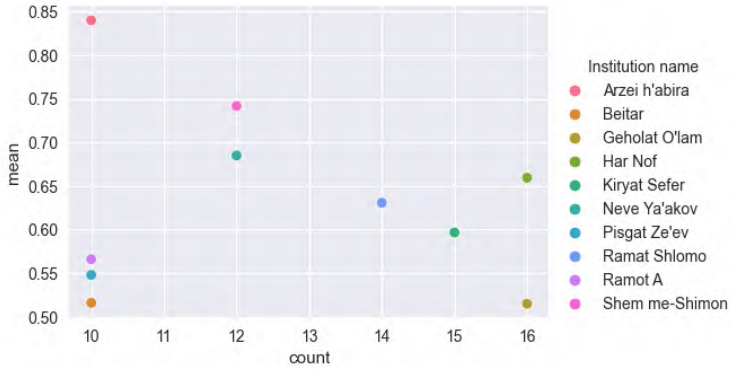


נתון שבולט מגרפים אלו הוא שבמוסד "ארזי הבירה" ממוצע המלגה גבוה וקבוע. היות שמוסד זה סומן כבר כאחד מהמוסדות שאופן חלוקת המלגות בו חריג ביחס לשאר המוסדות- מתחזקת ההשערה שאופן חלוקת המלגות ב"ארזי הבירה" שונה מהותית משאר המוסדות.



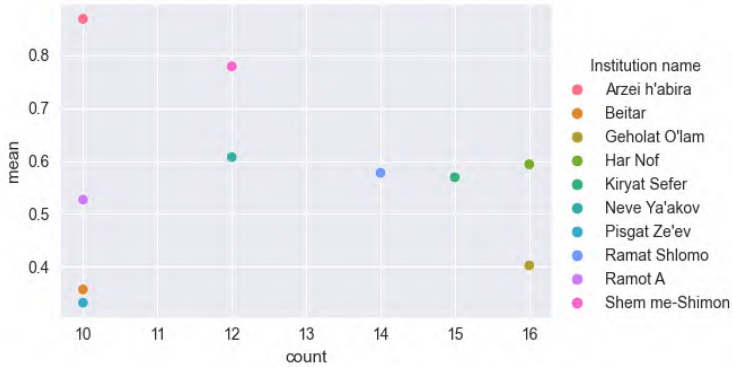
אבל, טרם נבדקה האפשרות שגם אחוז הנוכחות במוסד זה גבוה באופן יחסי, ולכן נבדוק את נתוני הנוכחות:

היחס בין אחוז הנוכחות לכמות הלומדים דומה מאוד ליחס שבין ממוצע המלגות לכמות הלומדים, מה שמאשש את הנחה שהוצבה כבר בשלב בדיקת היחס שקיים קשר הדוק בין אחוז הנוכחות לגובה המלגה.



הטענה שהצבנו קודם כי ב"ארזי הבירה" אופן חלוקת המלגות שונה- הופרכה, היות שגם אחוז הנוכחות הממוצע גבוה וקבוע.

אם כן, מסתבר שקיים נתון נוסף המשפיע על אופן חלוקת המלגות במוסדות שסומנו כחריגים, אך אינו מופיע בData שברשותנו.



בנוסף, ישנם 2 נתונים בעייתיים: מלגות גבוהות ללא קשר לאחוז הנוכחות, ואחוזי נוכחות ללא מלגות.

אציין, שאין בידי כלים לנתח ולדייק את נתונים אלו, ולכן ההמלצה שאתן היא ליצור נוהל אחיד לחלוקת המלגות, ומעקב אחריו. במידה ונהלי חלוקת מלגות לא יפתרו את בעיה זו- יהיה צורך לפנות אל בסיסי הנתונים ולהבין את חריגות אלו.

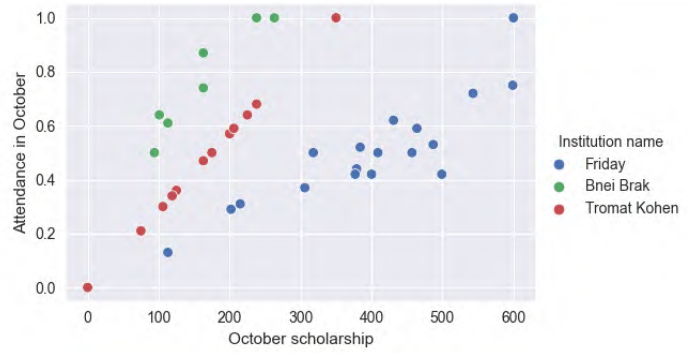
מסקנות:

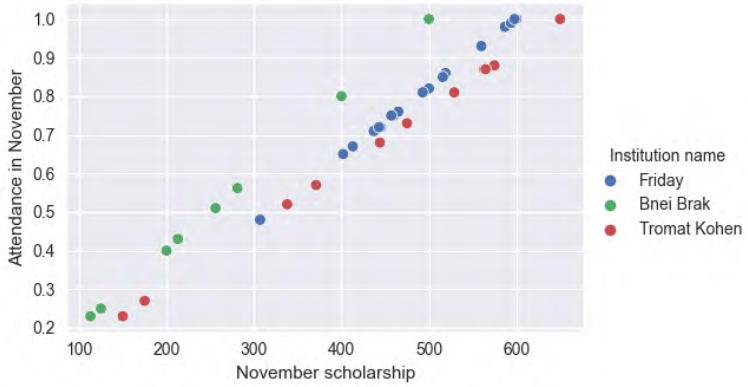
1. הData הקיים ברשותי במוסדות מסוג זה לא מדויק, כך שהמסקנות ניתנות בערבון מוגבל.
2. קיים יחס ישר בין אחוז הנוכחות לגובה המלגה. כאשר מ-90% נוכחות המלגה המתקבלת לרוב היא מלאה.

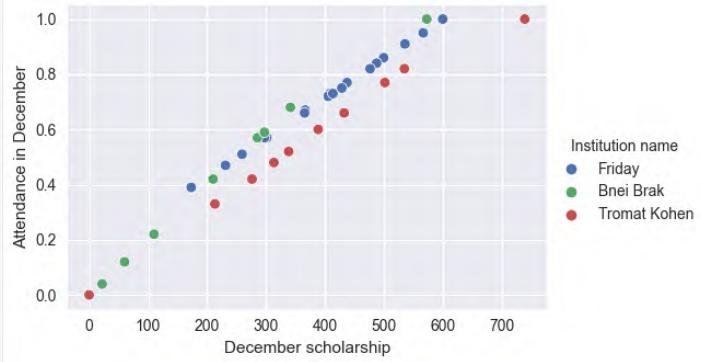
   במוסדות "ארזי הבירה" ו"שם משמעון" היחס בין אחוז הנוכחות לגובה המלגה שונה מהיחס במוסדות השונים בסיווג זה. מסתבר שיש נתון נוסף המשפיע על גובה המלגה שאינו קיים בData הנוכחי.
3. במקרים של חריגות קבוצתיות כמו מלחמה או חגים, אין נוהל אחיד לחישוב גובה המלגה, גם ללא קשר לאחוז הנוכחות. כך שיש ליצור נהלים אחידים למקרי קיצון, או נהלי קבלת החלטות במקרים אלו כדי ליצור אחידות.
4. יש לאסוף Data נוסף על מנת להבין מהן המלגות הגבוהות שניתנות ללא קשר לנוכחות, וכן מהם אחוזי הנוכחות ללא מלגה המשולמת עליהם כפי שפורט.

מוסדות סוף שבוע (Type 3):

ניתוח הקשר בין אחוז הנוכחות לגובה המלגה לפי החודשים השונים:



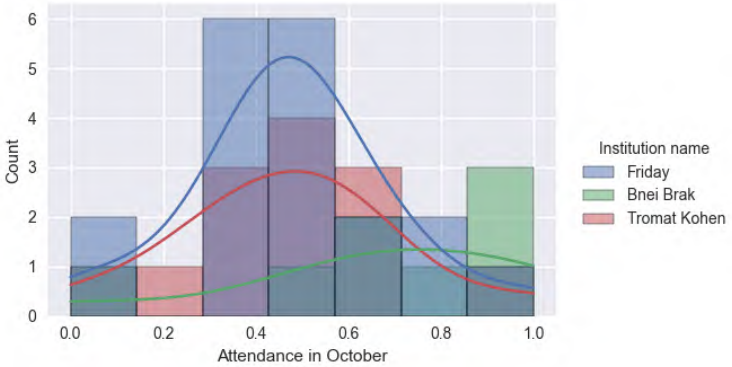




<u>תוצאות העולות מהגרפים:</u>

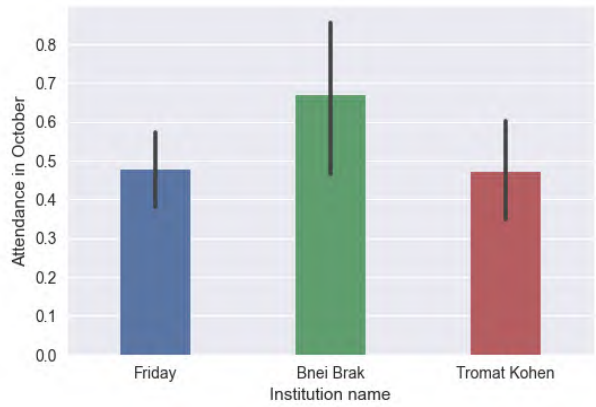
1. יש הלימה ברורה וחדה בין אחוזי הנוכחות לגובה המלגה.
2. היחס בין אחוז הנוכחות לגובה המלגה אינו אחיד בין המוסדות.
3. בחודש אוקטובר במוסדות "שישי" ו"בני ברק" אין קו ברור בחלוקת המלגה.
   בחודש זה גם גובה המלגה המלאה אינו אחיד בין המוסדות השונים.

כדי להבין את הסיבה לפיזור הנתונים בחודש אוקטובר, אציג את אחוזי הנוכחות וסכומי המלגות בחודש זה באופן מפורט על ידי תרשימים מסוג Histogram ו Bar:



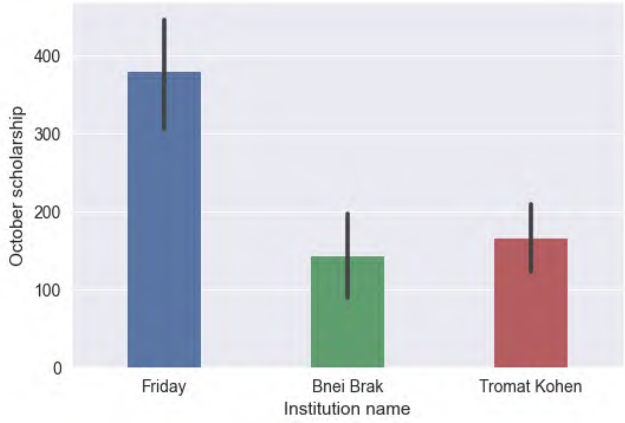
מעניין לגלות שאחוז הנוכחות מושפע מאזור האיום המלחמתי בחודש זה. כאשר בעיר בני ברק (בה היו אזעקות על פי נתוני פיקוד העורף בכל סוף שבוע בחודש אוקטובר) רמת הנוכחות המצטברת נמוכה ביחס למוסדות "שישי" ו"תרומת כהן" המתנהלים בעיר ירושלים (בה היו אזעקות רק בסוף השבוע הראשון בחודש אוקטובר).

אך עדיין בהצגת הנתונים במבט על ממוצעי נוכחות התקבלו תוצאות מפתיעות:



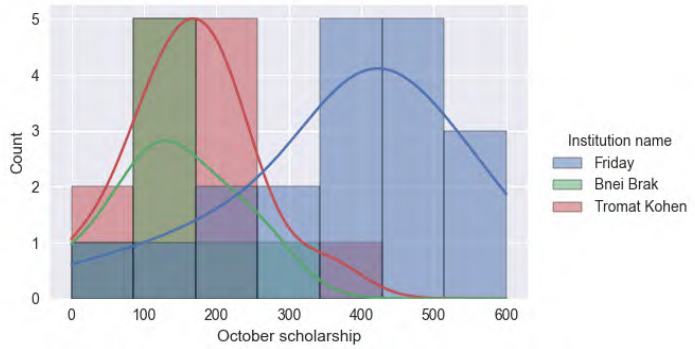
ממוצע הנוכחות ללומד בבני ברק- בה הכמות המצטברת של הנוכחים אינה גבוהה- גבוה באופן משמעותי מהמוסדות האחרים מסוג זה.

אבל:



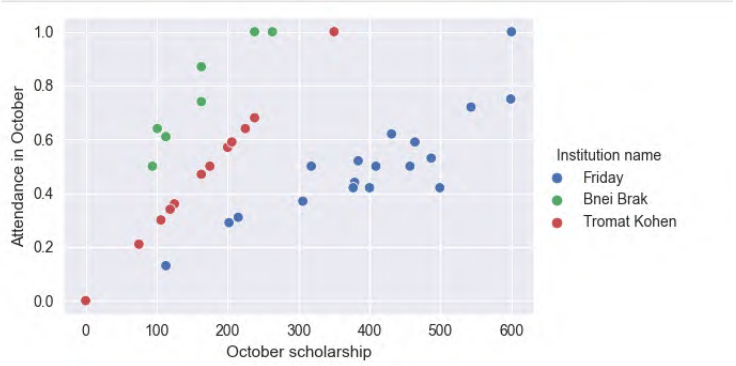
ממוצע המלגות ללומד שונה משמעותית מממוצע הנוכחות באותו המוסד!.

כדי לדייק את המסקנות העולות, הצגתי גם את הסכומים המצטברים של המלגות לפי מוסדות:



ההתפלגות בסכומי המלגות וקו המגמה שנוצר- שונים באופן ניכר מההתפלגות וקו המגמה ברמת הנוכחות באותו חודש.

גם מלמידת התרשים הScatter שיצרתי בתחילת התהליך ניתן לראות:



לומדי המוסד "שישי" קבלו מלגות גבוהות ביחס לאחוז הנוכחות, בעוד לומדי מוסד "בני ברק" קבלו מלגות נמוכות ביחס לאחוז הנוכחות. יש לציין כי במוסד "תרומת כהן" שמרו על היחס הישר בין אחוז הנוכחות לגובה המלגה.

בשל נתונים חריגים אלו, ששונים באופן ניכר מנתוני החודשים הבאים, מסקנתי היא שאין להקיש בין נתוני חודש זה לנתוני שאר חודשי הרבעון.

הנחת היסוד בסיטואציה זו היא שמצב המשק וחגי החודש יצרו בכל מוסד שיטת חישוב שונה, שהובילה לתוצאות שונות.

נתונים אלו מובאים ומודגשים, על מנת שבעת יצירת נהלי חלוקת מלגות לסוג מוסדות זה, יועלו נהלי חלוקת מלגות ברורים למצבי קיצון שונים שאינם צפויים מראש.

כעת, לאחר שבודדתי את נתוני חודש אוקטובר, אמשיך ללמוד את נתוני מלגות חודשי נובמבר- דצמבר:

נתוני ממוצע בחלוקה לפי חודש ומוסד שהסקתי בטבלה:

```
t3.groupby('Institution name')[['October scholarship','November scholarship','December scholarship']].mean()
```

Institution name	October scholarship	November scholarship	December scholarship
Bnei Brak	141.875000	261.000000	237.125000
Friday	378.052632	486.052632	412.894737
Tromat Kohen	165.000000	437.000000	333.384615

```
t3.groupby('Institution name')[['Attendance in October','Attendance in November','Attendance in December']].mean()
```

Institution name	Attendance in October	Attendance in November	Attendance in December
Bnei Brak	0.670000	0.522750	0.455000
Friday	0.475263	0.797895	0.732632
Tromat Kohen	0.471538	0.671538	0.502308

ניתן לראות כי ממוצע המלגות וממוצע הנוכחות במוסד "בני ברק" נמוך מממוצע כלל המוסדות.

בעוד במוסד "שישי" הנתונים גבוהים מהממוצע הכללי ומנתוני המוסדות האחרים.

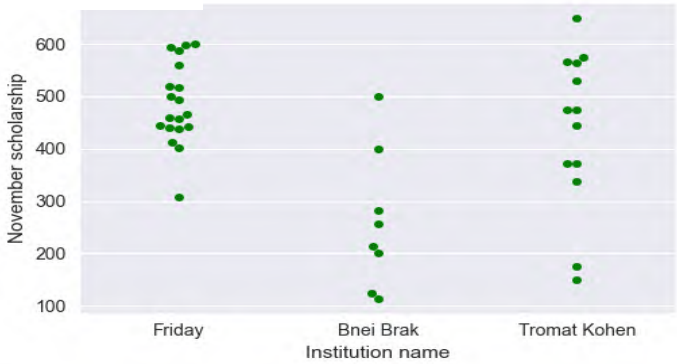
```
t3[['October scholarship','November scholarship','December scholarship']].mean().to_frame()
```

	0
October scholarship	261.575
November scholarship	425.100
December scholarship	351.900

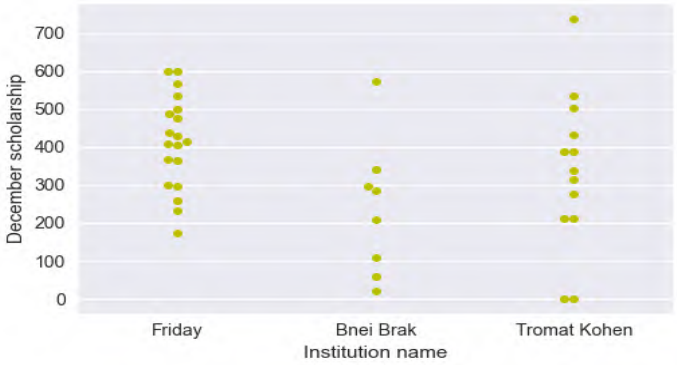
```
t3[['Attendance in October','Attendance in November','Attendance in December']].mean().to_frame()
```

	0
Attendance in October	0.51300
Attendance in November	0.70180
Attendance in December	0.60225

ובאופן גרפי:

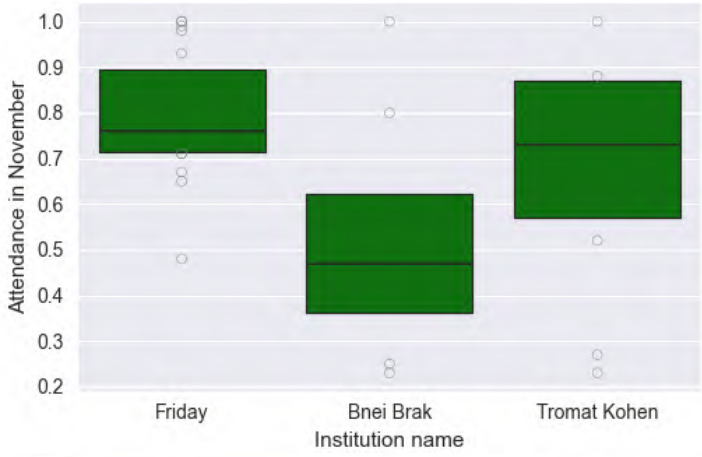


פיזור סכומי המלגות מלמד כי לומדי המוסד "בני ברק" מקבלים מלגות נמוכות יותר ביחס ללומדי מוסדות אחרים מסוג זה.

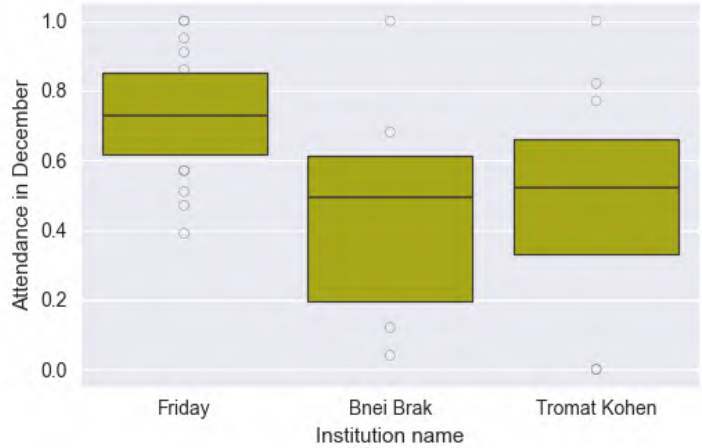


כדי להבין את נתון זה יש לבדוק האם באופן יחסי לומדי המוסד "בני ברק" נוכחים פחות מלומדי מוסדות אחרים בסיווג זה:

כדי להבדיל בתצוגה בין נתוני המלגות לנתוני הנוכחות, הגדרתי את תרשימי הנוכחות באופן שונה בו מודגש גם החציון ופיזור הנקודות בהשוואה לכך:



המידע העולה הוא כי אחוזי הנוכחות הממוצעים והתפלגותם במוסד "בני ברק" אכן נמוכים מאחוזי הנוכחות, התפלגותם וריכוזם במוסדות אחרים.



בחודש דצמבר יש צמצום פערים בין המוסדות השונים הן מבחינת אחוזי הנוכחות והן מבחינת סכומי המלגות שחולקו.

מסקנות:

1. יש הלימה ברורה בין אחוזי הנוכחות לגובה המלגות, ללא חריגים מיוחדים.
2. הקשר בין אחוז הנוכחות לגובה המלגה אינו אחיד בכלל המוסדות בסיווג זה, מה שאומר שיש לנסח נוהל אחיד המקשר בין אחוז חיסורי הנוכחות לשיעור הפחתת גובה המלגה.

יש לציין כי יתכן שהשוני בגובה המלגות נובע מתשלומי בונוס נוספים, כדוגמת תשלום על מבחן או רציפות בלימוד, היות שנתוני הData שבידי לא מתחשבים בנתונים חריגים- מסקנה זו תעבור בדיקת אמיתות מול מנהלי המוסדות בפועל.

3. במקרים של חריגות קבוצתיות כמו מלחמה או חגים, אין נוהל אחיד לחישוב גובה המלגה, גם ללא קשר לאחוז הנוכחות. כך שיש ליצור נהלים אחידים למקרי קיצון, או נהלי קבלת החלטות במקרים אלו כדי ליצור אחידות.