




תרגיל מסכם

Big Data

פברואר 2019

רון לוי



# exploratory analysis

## הגדרות המשתנים

number of pickups - ספירת ההזמנות מ-uber בזמן נתון

wDay - האם מדובר ביום חול או בסוף-שבוע

bikes pickups - ספירת השכרות האופניים בזמן נתון

Rain - האם ירד גשם או לא ירד גשם. 0 אומר שלא ירד גשם, 1 אומר שירד.

## בכדי לארגן את הנתונים, ביצעתי את הפעולות הבאות:

1. השארתי רק תצפיות מהרדיוס הרלוונטי
2. חילקתי את הדאטה לאינטרוולים של רבע שעה
3. מיזגתי את החודשים יולי, אוגוסט וספטמבר (השבועיים הראשונים)
4. הבדלתי בין ימי חול לסופי-שבוע, והגדרתי חגים כסופי-שבוע
5. הוספתי דאטה של השכרות אופניים ושל תנועת הרכבות

## summary של הדאטה

number_of_pickups	wDay	bikes_pickups
Min. : 1.00	weekday:5144	Min. : 1.00
1st Qu.: 11.00	weekend:2298	1st Qu.: 13.00
Median : 22.00	NA	Median : 44.00
Mean : 24.98	NA	Mean : 47.47
3rd Qu.: 35.00	NA	3rd Qu.: 68.00
Max. :105.00	NA	Max. :244.00

## כמות הימים הגשומים

ירד גשם	לא ירד גשם
<b>23,821</b>	<b>162,064</b>

## מה הקשר של המשתנים שבחרתי למשתנה התלוי?

wDay - יום עבודה מול יום חופש

בתחילת העבודה העליתי השערה כי יש קשר בין אופי היום לבין הביקוש ל-uber. האינטואיציה שלי אמרה שאנשים משתמשים ב-uber לצרכי היום-יום, וממעיטים בשימוש בימי מנוחה.

bikes pickups - השכרות אופניים

אחת ההשערות שרציתי לבדוק, היא האם יש קשר בין השימוש ב-uber לבין השימוש באופניים שכורות. הנחתי כי ישנו קשר, גם אם לא סיבתי, בין שני המשתנים. כאשר בחנתי את הקורלציה ביניהם הבנתי כי ייתכן וזו תהיה תוספת טובה למודל.

## number of pickups לבין bikes pickups

0.59

Rain

רציתי לבדוק האם יש השפעה לכך שיורד גשם על כמות ההזמנות מ-uber.

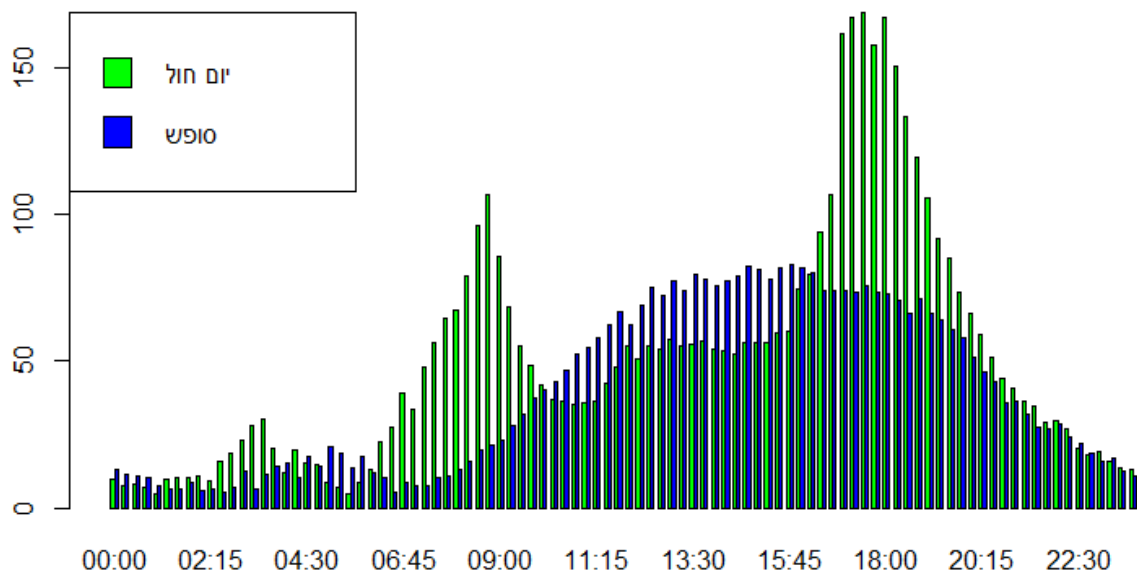
## נבדוק - האם קיים הבדל בשימוש ב-uber ביום חול ובסופ"ש?

ביצעתי t-test לרגרסיה שבה השימוש ב-uber הוא המשתנה התלוי, ו-wDay הוא המשתנה הבלתי תלוי.

קיבלתי שה-  $p\text{-value} < 2.2e-16$ , כלומר:

**קיים הבדל מובהק!**

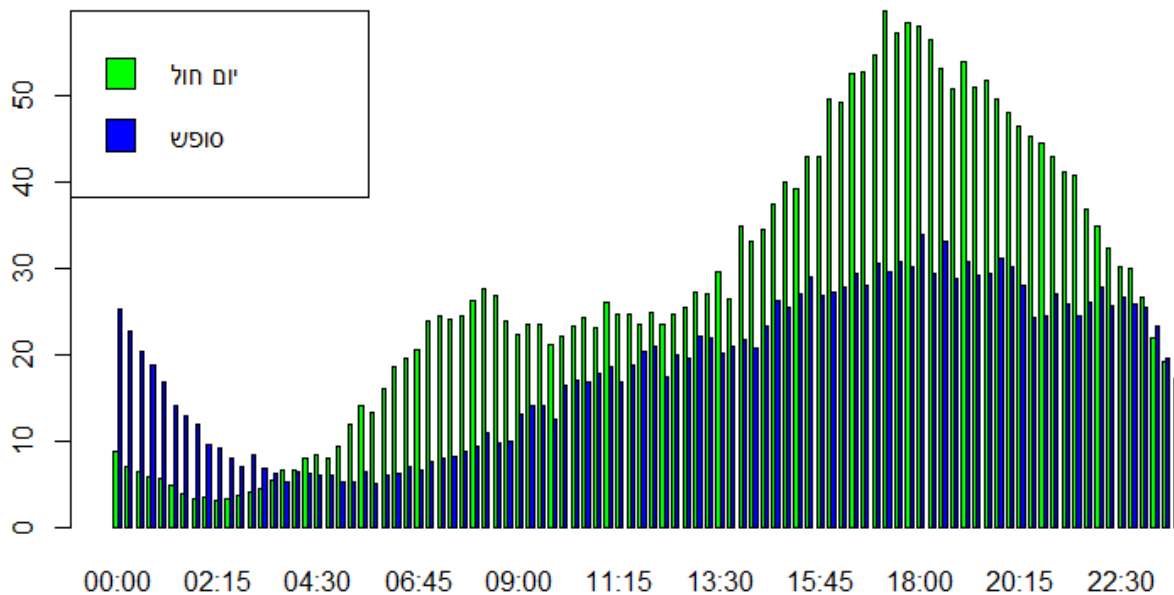
### השכרות אופניים לפי שעה



ניתן לראות את השוני בהשכרת האופניים בין ימי חול לסופי-שבוע. השימוש בימי חול עולה בצורה חדה החל מ-6:00, ומגיע לשיא בשעה 9:00 שעת תחילת העבודה. השיא הנוסף הוא בשעה 18:00, אז מרבית האנשים עוזבים את המשרד.

**השימוש בסופי-שבוע** שונה מאשר בימי חול. הביקוש עולה בצורה מתונה לאורך הבוקר, ומגיע לשיא בסביבות 15:00. הביקוש בסופי-שבוע הרבה יותר "חלק", אין שינויים דרסטיים כמו בימי חול.

### שימוש לפי שעה - uber



גם כאן ישנו הבדל די ברור בין הימים השונים. השימוש בימי חול עולה החל מ4:30 בבוקר, ומגיע לשיא בשעה 9:00. כמו אצל השימוש באופניים, גם כאן ניתן לראות עלייה חדה בשימוש בשעה 18:00, שעת סיום העבודה. במקרה של uber, העלייה אף יותר קיצונית.

השימוש בסופי-שבוע שונה מביום חול. ישנו ביקוש ל-uber גם באמצע הלילה, והוא הולך ונחלש לקראת הבוקר. הביקוש מגיע למינימום בשעה 6:00 לערך, ואז עולה בצורה מתונה עד השעה 18:00. בניגוד ליום חול ובדומה לאופניים, העלייה הרבה יותר מתונה.

### מסקנות

1. ככל הנראה, אנשים משתמשים ב-uber בעיקר למטרות עבודה
2. ככל הנראה, אנשים משכירים אופניים בעיקר למטרות עבודה
3. גם אם אנשים אינם משתמשים ב-uber ומשכירים אופניים בעיקר למטרות עבודה, נוכל להניח כי סיבה משותפת אחרת להרגלי הצריכה של uber ושל האופניים; זאת מכיוון שההתפלגות של שני המשתנים נראית דומה
4. יש הבדל די משמעותי בין יום חול לסוף-שבוע
5. ל-uber יש ביקוש גם בשעות מאוחרות בלילה, אך בעיקר בסופי שבוע. ככל הנראה בגלל בלוינים
6. על כל יום גשום ישנם כשבעה ימים ללא גשם (בתקופה המדוברת על-פי מדגם זה)

## תיאור המודל שנבחר

בכדי לבחון את מידת הדיוק של המודלים השתמשתי בהגדרה הבאה:

$$ACCURACY = 1 - \left( \frac{RMSE}{Mean\ Hourly\ Demand} \right)$$

המודל הראשון אותו בחנתי הוא:

$$uber_{count} = Hour + wDay + bikes_{count} + Rain$$

רמת הדיוק שלו היא:

0.5635747

המודל השני שבדקתי:

$$uber_{count} = Hour * wDay + Hour + wDay + bikes_{count} + Rain$$

רמת הדיוק שלו היא:

0.5987158

השלישי:

$$\log(uber_{count}) = +Hour + wDay + bikes_{count} + Rain$$

לאחר שקיבלתי את התחזית, הפעלתי אקספוננט. רמת הדיוק:

0.561969

הרביעי:

$$\log(uber_{count}) = Hour * wDay + Hour + wDay + bikes_{count} + Rain$$

רמת הדיוק:

0.5881397

נבדוק מודל בלי משתנה האופניים:

$$uber_{count} = Hour * wDay + Hour + wDay + Rain$$

רמת הדיוק:

0.5968815

### קיבלתי שהמודל השני הגיעה לדיוק המרבי, ולכן אבחר בו.

עקב הגבלת העמודים בעבודה זו, לא יכולתי להוסיף את ערכי המקדמים עבור המודל הנבחר. אציין כי המשתנים "Rain", "bikes\_pickups", "wDay" הינם מובהקים. האינטראקציה בין "weekend" לבין "hour" הינה מובהקת, מלבד בשעות הלילה המאוחרות (בין 0:00 ל-02:00). "hour" מובהק אחרי השעה 06:00.

נתונים נוספים לגבי הרגרסיה:

$$Adjusted R^2 = 0.7072$$

$$P - value < 2.2e-16$$