כלכלה – BIG DATA:

1. שאלה 1:

a. ניקוי והגמשת data:

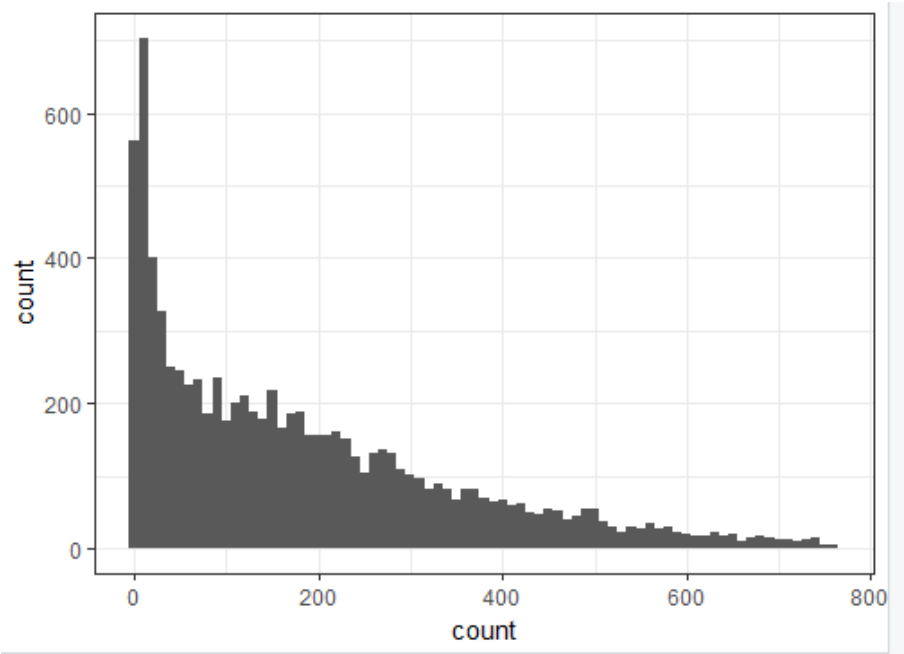
- לקחנו את כל ה-data עד ה-quantile ה-99.
- בדקנו שאין ערכי N/A ב-DataSet.
- יצרנו featureים חדשים של שעה (hour), יום בשבוע (weekday), זמן (time), תאריך (date), חודש (month) ושנה (year) המבוססים על ה-datetime.
- הפכנו משתנים לקטגוריאלים (ממספר לשם)

b. מסקנות כלליות:

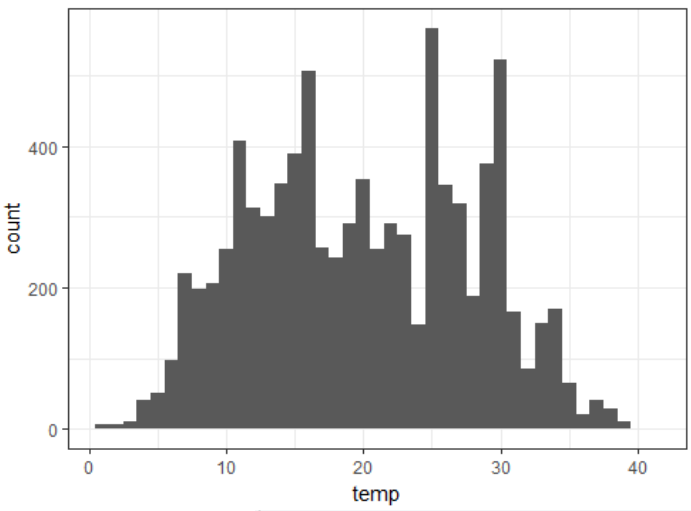
כמות ההשכרות של האופניים נעה בטווח רחב של בין השכרה אחת ביום לבין כ-760 כאשר הממוצע הוא 183 ליום:

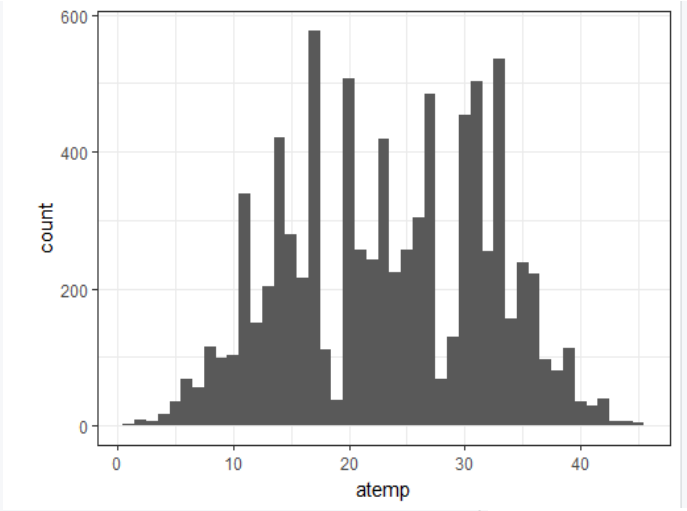
```
summary(bike_train_clean$count)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.0    41.0   141.0   183.9   277.0   761.0
```

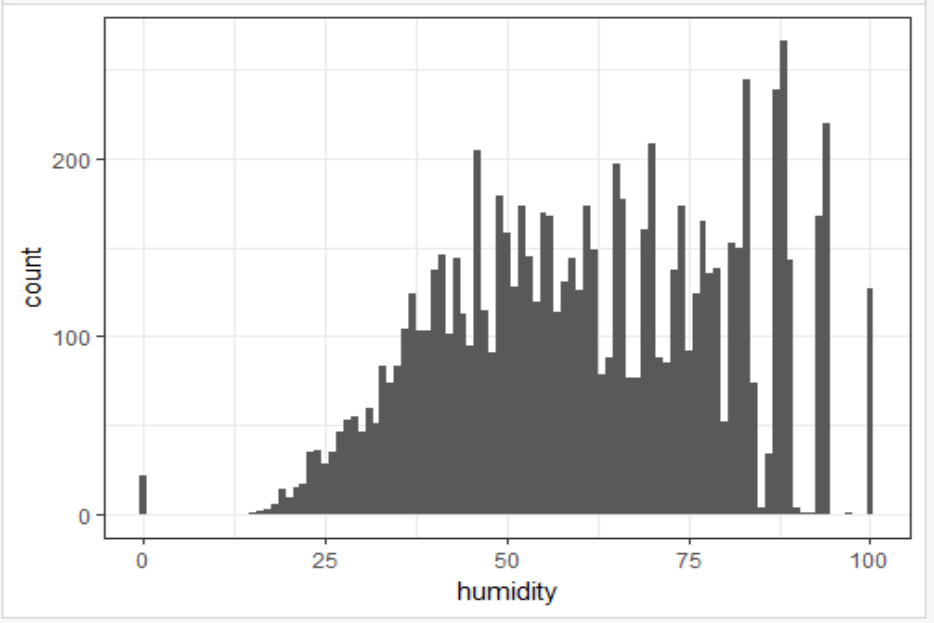
ניתן לראות שבאופן כללי feature ה- count מתנהג כמו רשת חברתית, זאת לפי ההיסטוגרמה שלו:



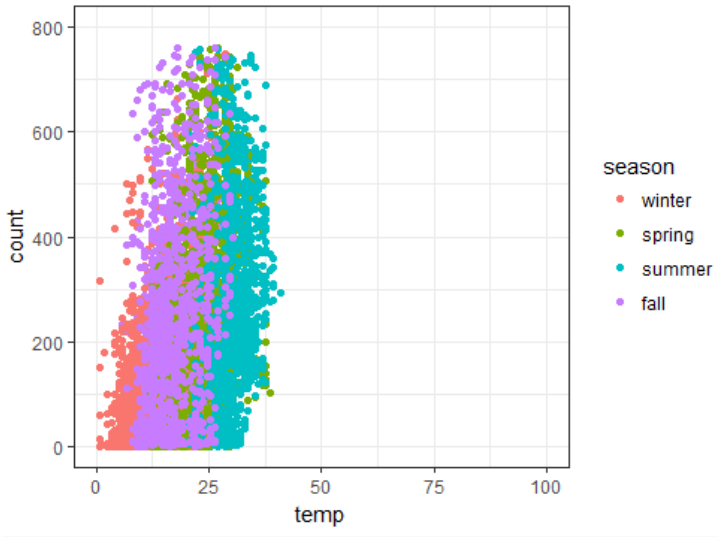
מבחינת טמפורטורה, לחות וטמפטורה מורגשת (temp, humidity, atemp) ניכר שאין הם משפיעים רבות על כמות אירועי ההשכרות מלבד בטמפרטורות הקיצוניות. כמו-כן לא ניתן ללמוד מכך רבות אלא רק לצפות במגמה כי מדובר בכמות אירועי ההשכרות במצבים הללו ולא כמה השכרות בוצעו בפועל במהלך תצפית (אירוע השכרה):

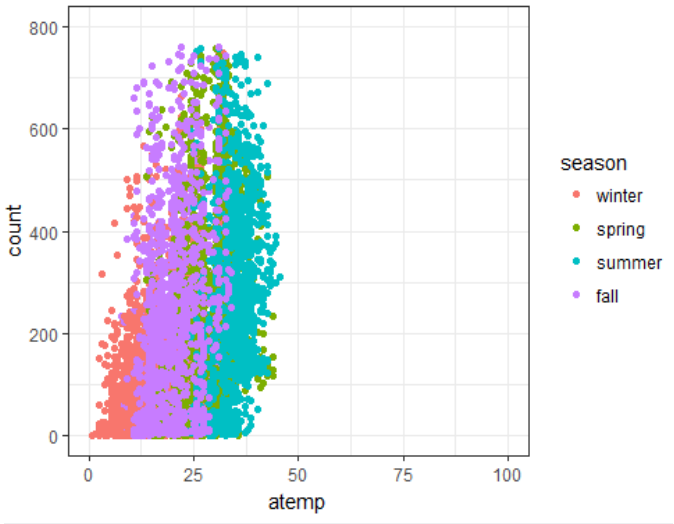


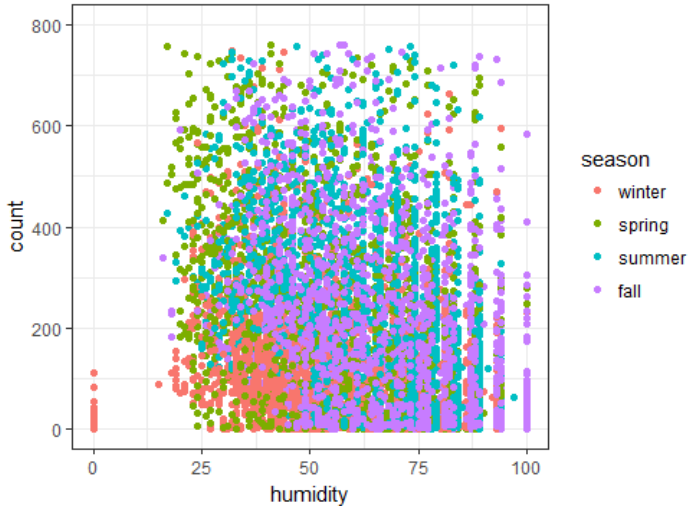




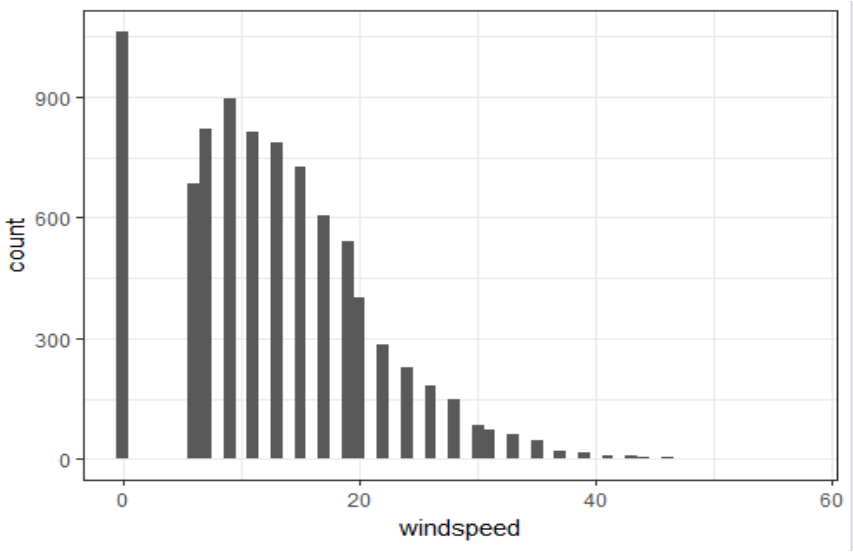
ניתן לראות שלמרות כי היינו חושבים שהלחות והמטפורות ישפיעו מאוד, הדבר אינו תואם למציאות:



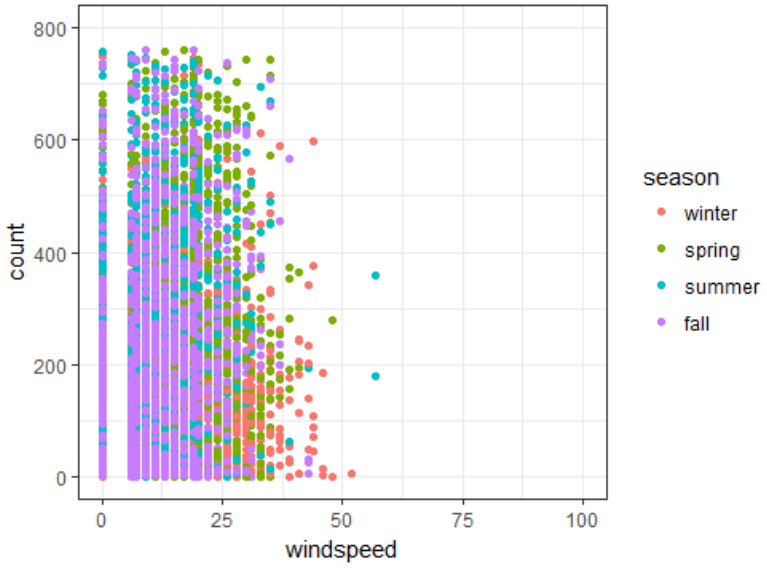




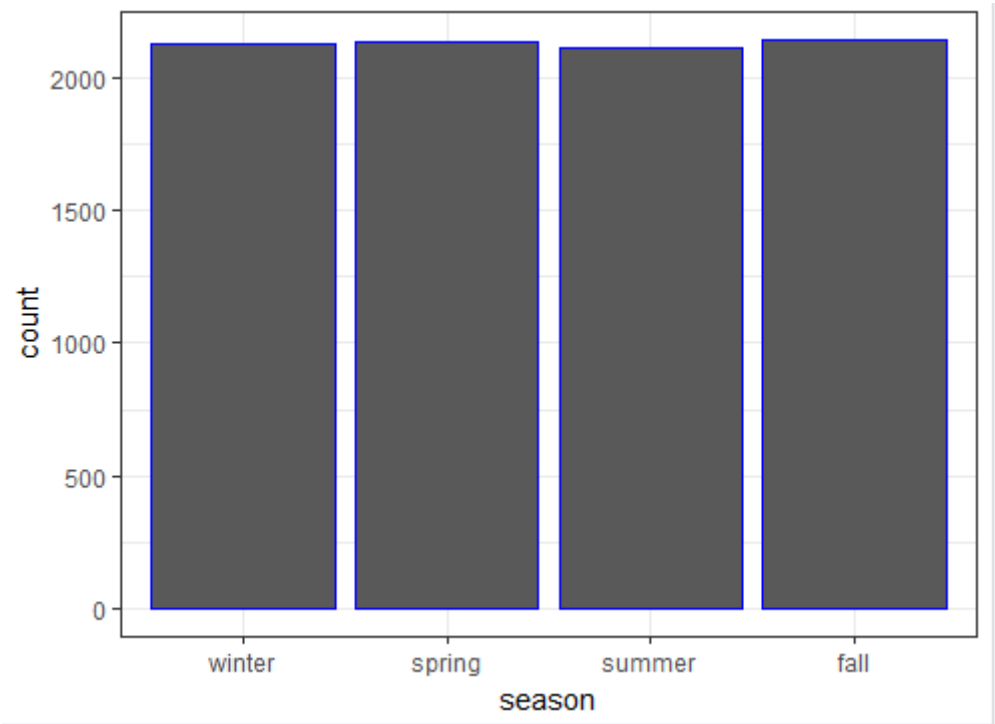
לעומת זאת, כן ניתן לראות במגמה כזאת מבחינת מהירות הרוח (windspeed) והיא שככל שהרוח חלשה ישנן יותר השכרות:



ניתן לראות גם מבחינת מהירות הרוח ביחס לעונות – ככה שהרוח מתחזקת ישנם פחות ופחות אירועי השכרות.



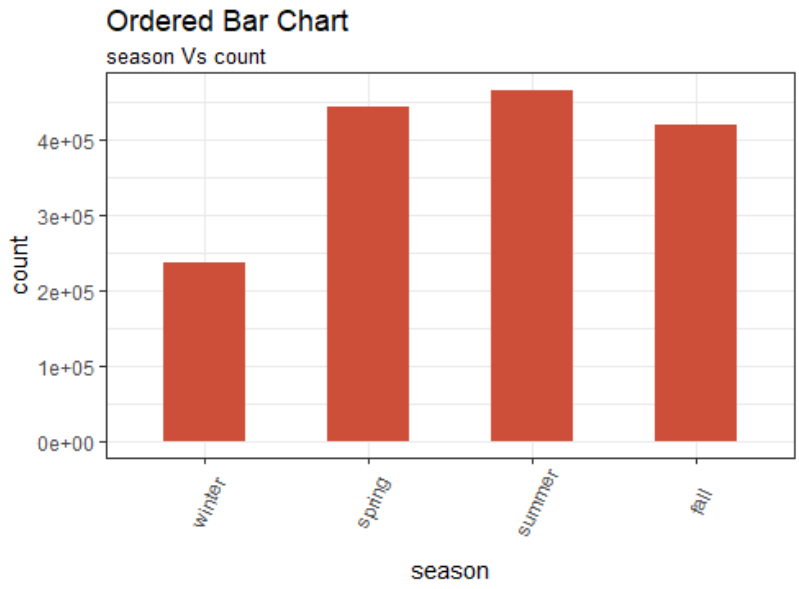
מבחינת משתנים קטגוריאלים, נראה כי מבחינת אירועי השכרות, לעונה אין השפעה :



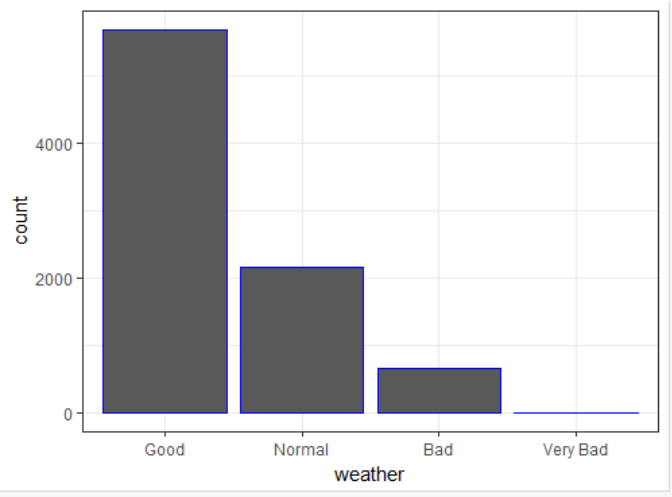
אך כאשר בודקים את הממוצע פר עונה ניתן לראות כי יש העדפה ברורה לקיץ, לאחר מכן לסתיו ולבסוף הכי פחות בחורף.

```
  season    count
1 winter 111.1990
2 spring 207.9925
3 summer 220.7136
4   fall 195.9178
>
```

נראה זאת באמצעות bar plot על הכמות :



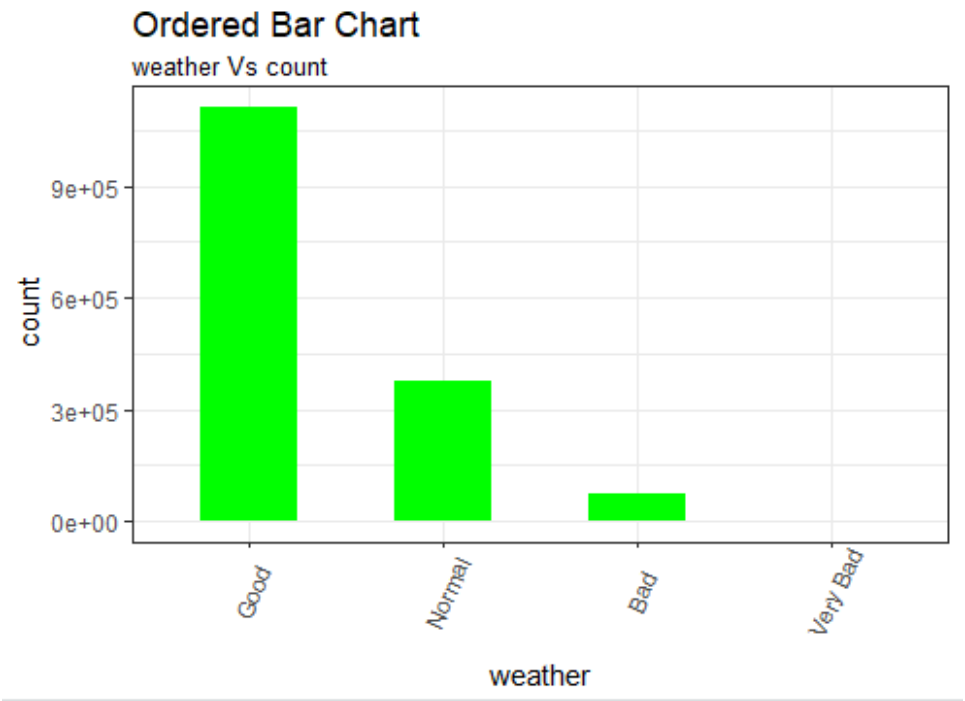
אולם מבחינת מזג אוויר (**weather**) נראה כי באופן מגמתי יש יותר אירועי השכרות כאשר מזג האוויר **נחשב טוב**:



ניתן לראות כי בממוצע אכן הדבר מתקיים, מלבד כך שבמזג אוויר רע יש פחות ממזג אוויר טוב:

```
   weather     count
1     Good 195.9348
2   Normal 174.6589
3      Bad 112.0641
4 Very Bad 164.0000
```

מבחינת כמויות בסה"כ לפי מזג האוויר – אכן הדבר זהה ל-avg. נראה זאת באמצעות bar-plot:



כאשר מסתכלים על יום עבודה לעומת יום חופשה (סופ״ש או חג) מקבלים כי ישנם יותר אירועי השכרה ב- **workingday** :

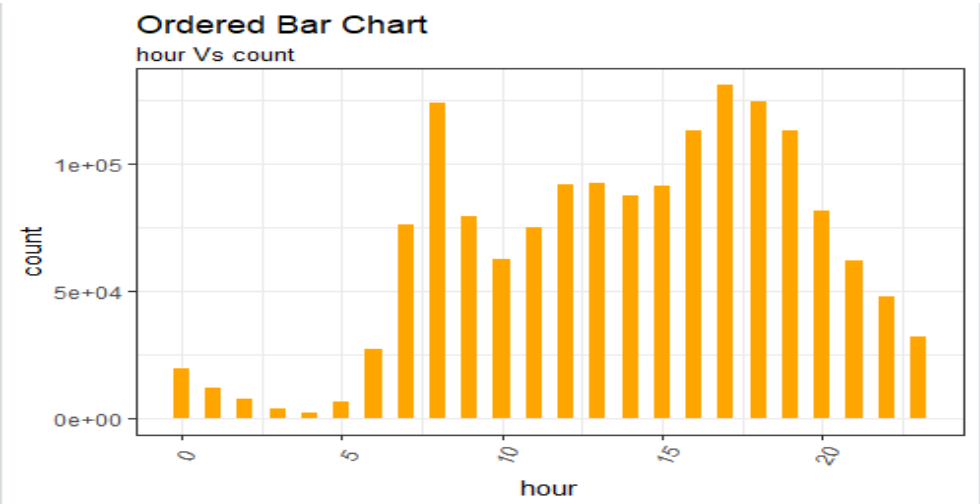
אולם ישנם יותר ימי עבודה מימי חופשה. לכן, עפ״י ממוצע, ניתן לראות כי מספר אירועי ההשכרה **דומה** :

```
      workingday    count
1 not workingday 187.8487
2     workingday 182.0877
>
```

כעת, נדון על סטטיסטיקות של זמנים בהתאם לכמויות. נחלק זאת ל-3 משתנים עיקריים שבהם נדון - חודש (month), יום בשבוע (weekday) ושעה (hour).

<u>שעה</u>:

נראה כי השעות בהן ישנן הכי הרבה השכרות אופניים הם 08:00 בבוקר, 17:00 ו-18:00 :



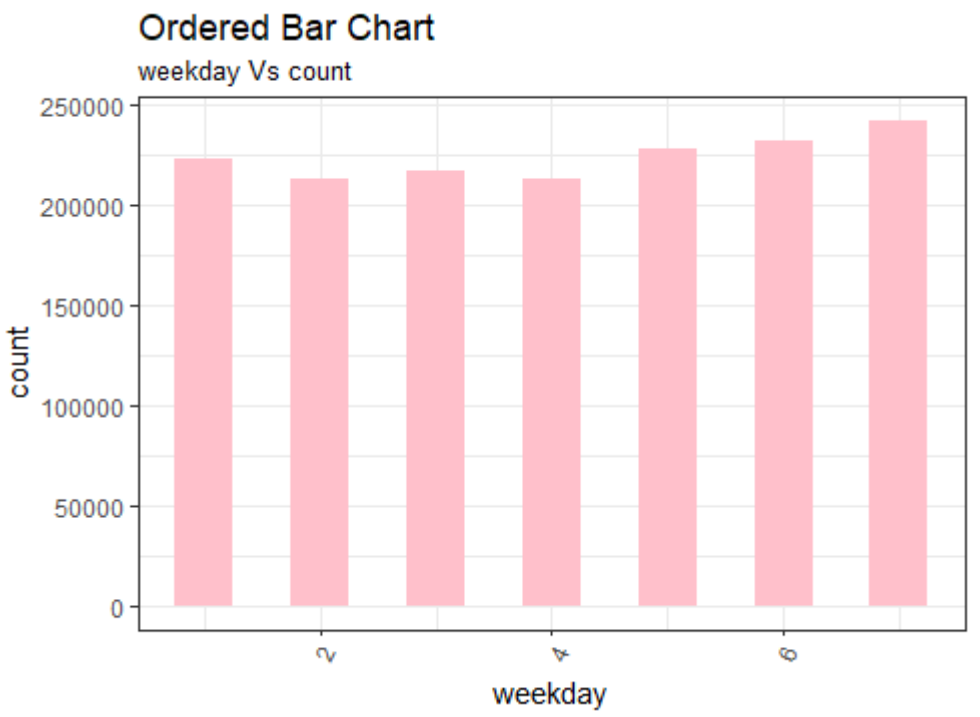
הדבר נתמך גם באופן **ממוצע**:

	hour	count
1	0	55.094444
2	1	34.041783
3	2	22.580282
4	3	11.847507
5	4	6.367816
6	5	19.450980
7	6	75.430556
8	7	210.822222
9	8	351.059490
10	9	220.066667
11	10	173.355556
12	11	208.622222
13	12	255.897222
14	13	257.341667
15	14	243.272222
16	15	253.886111
17	16	314.749304
18	17	414.056962
19	18	383.541538
20	19	313.669444
21	20	226.486111
22	21	172.494444
23	22	132.619444
24	23	89.788889

> |

<u>יום בשבוע:</u>

נראה כי באופן יחסי, בסופ״ש ישנן יותר השכרות אופניים בסה״כ:

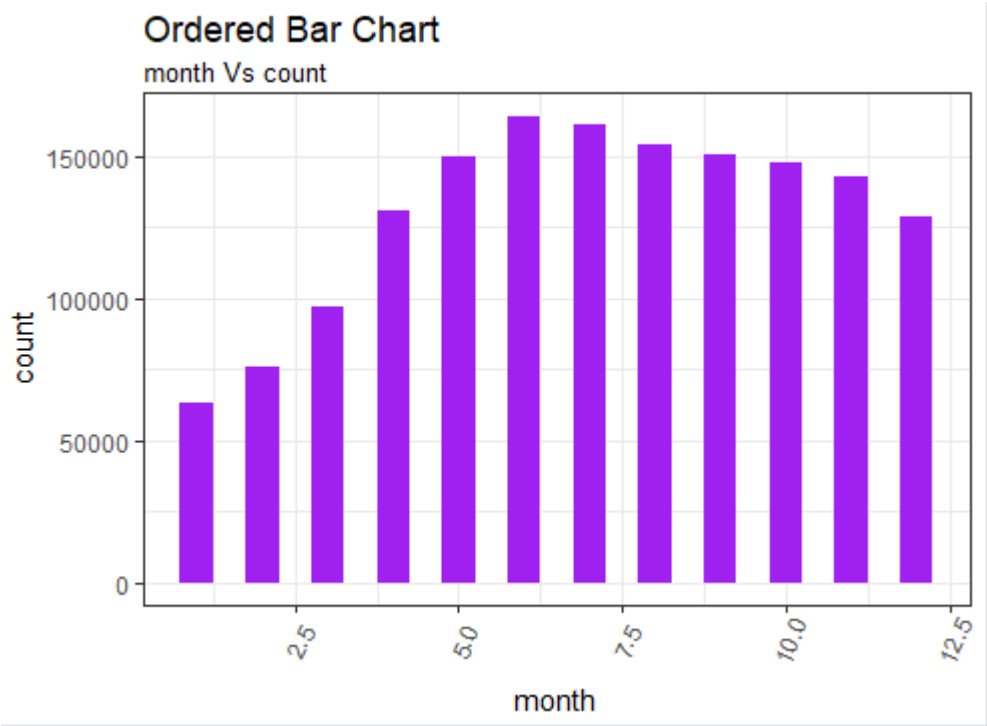


אולם מבחינת **ממוצע**, הדבר הינו קצת **שונה** ובעיקר בשישי-שבת יש כ-15 יותר השכרות:

```
  weekday    count
1       1 179.3323
2       2 181.1770
3       3 177.1012
4       4 177.8545
5       5 186.2424
6       6 191.8632
7       7 193.7362
>
```

<u>חודש:</u>

מבחינת חודשים נראה כי חודשי הקיץ הינם החודשים בהם יש הכי הרבה השכרות:



הדבר נתמך גם מבחינת ממוצע בחודשים הללו לאירועי השכרה:

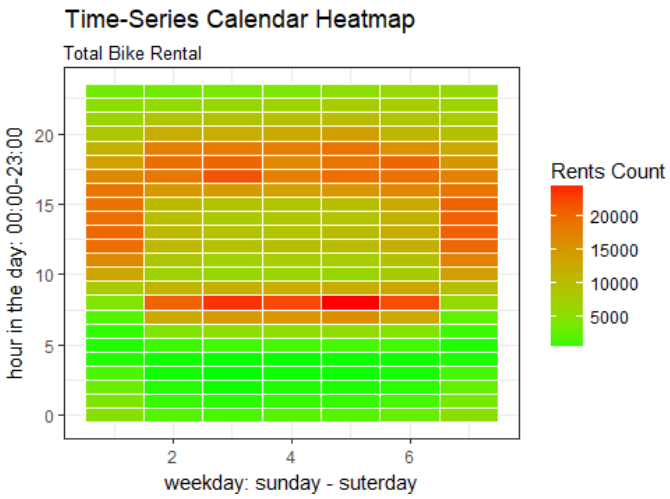
	month	count
1	1	89.86686
2	2	106.75352
3	3	136.85634
4	4	182.67832
5	5	209.16760
6	6	232.43768
7	7	226.38819
8	8	219.22080
9	9	216.42241
10	10	210.29202
11	11	198.91516
12	12	178.90972

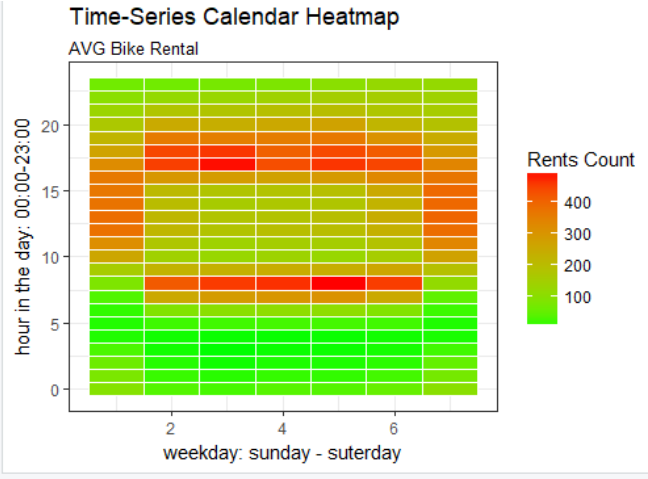
> |

כעת נראה מפות חום הממחישות את הנתונים הללו.

לכל זוג נתונים נחקר (יחד עם נתון ה-count בתור נתון שלישי), נראה מבחינת **ממוצע** (ימין) ומבחינת **כמות** השכרות כוללת (שמאל):

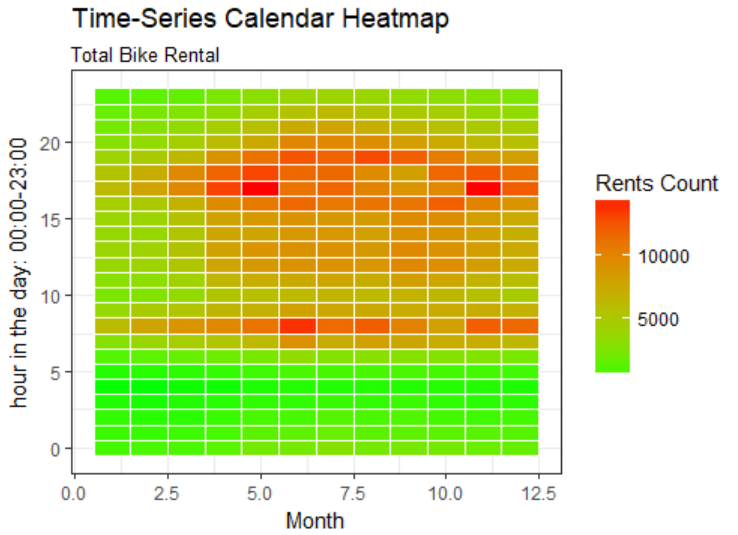
<u>יום VS שעה VS כמות:</u>

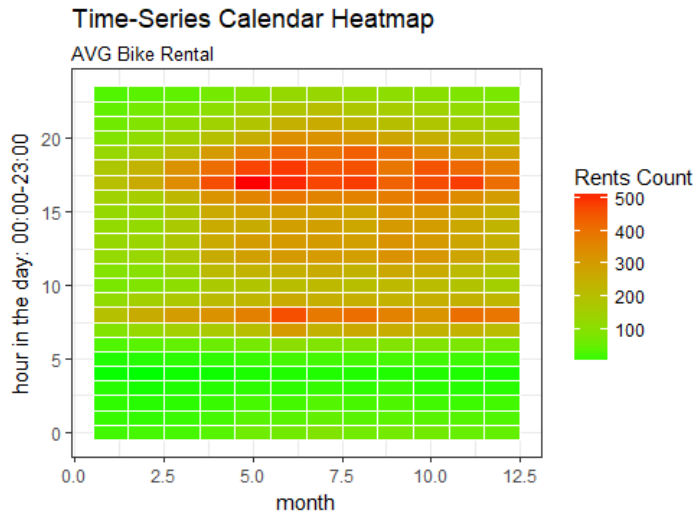




ניתן לראות בצורה נורא יפה, כי בסופ״ש (שבת וראשון) רב ההשכרות מתרחשות יותר בזמנים של בוקר מאוחר-צהרים ואילו בימות השבוע, רב ההשכרות מבוצעות בבקרים ואחה״צ.

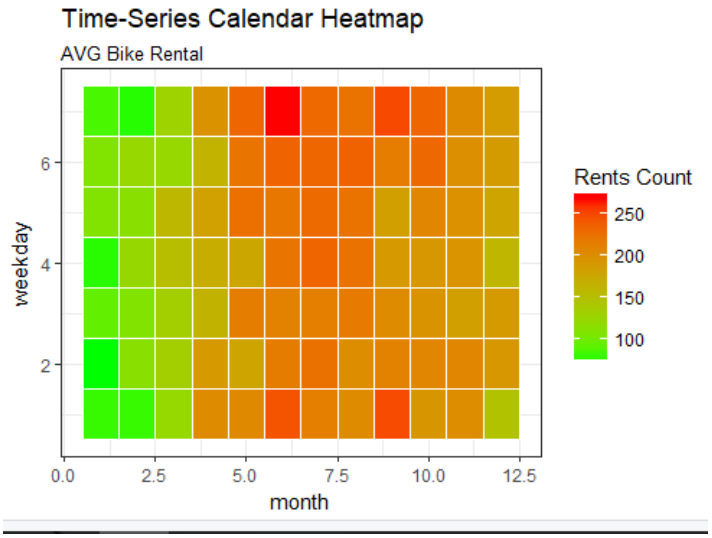
<u>שעה VS חודש VS כמות:</u>

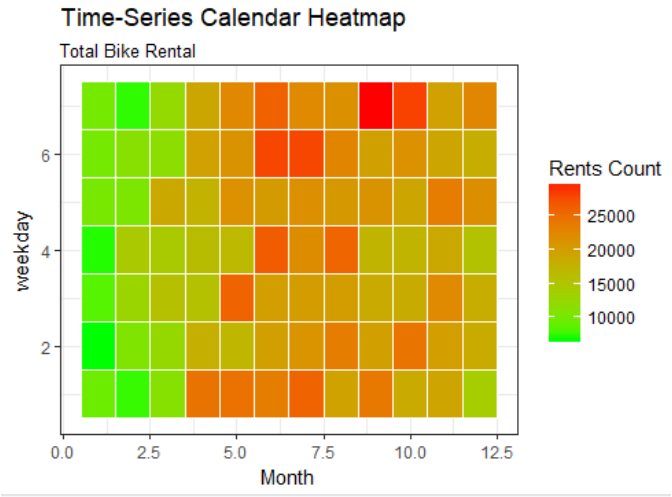




ניתן לראות כי אכן חודשי הקיץ הינם חודשים מובילים כצפוי.

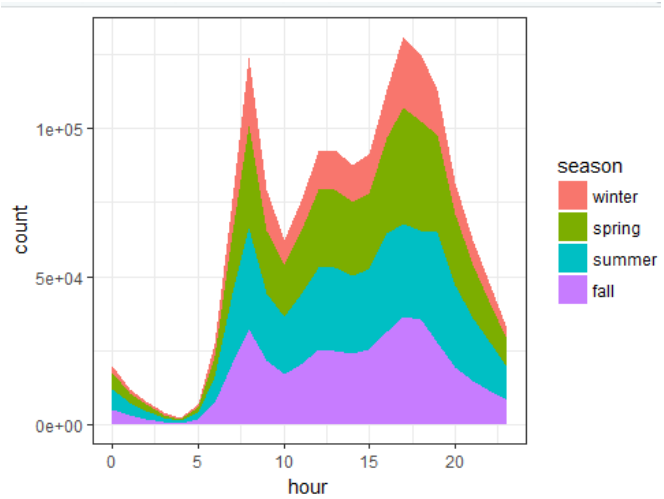
יום VS חודש VS כמות:

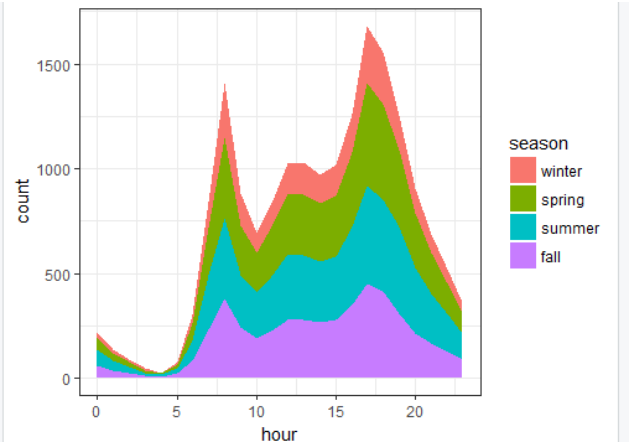




ניתן לראות כי הזמנים המשמעותיים ביותר להשכרות הינם בקיץ ובסופ״שים, כמקודם.

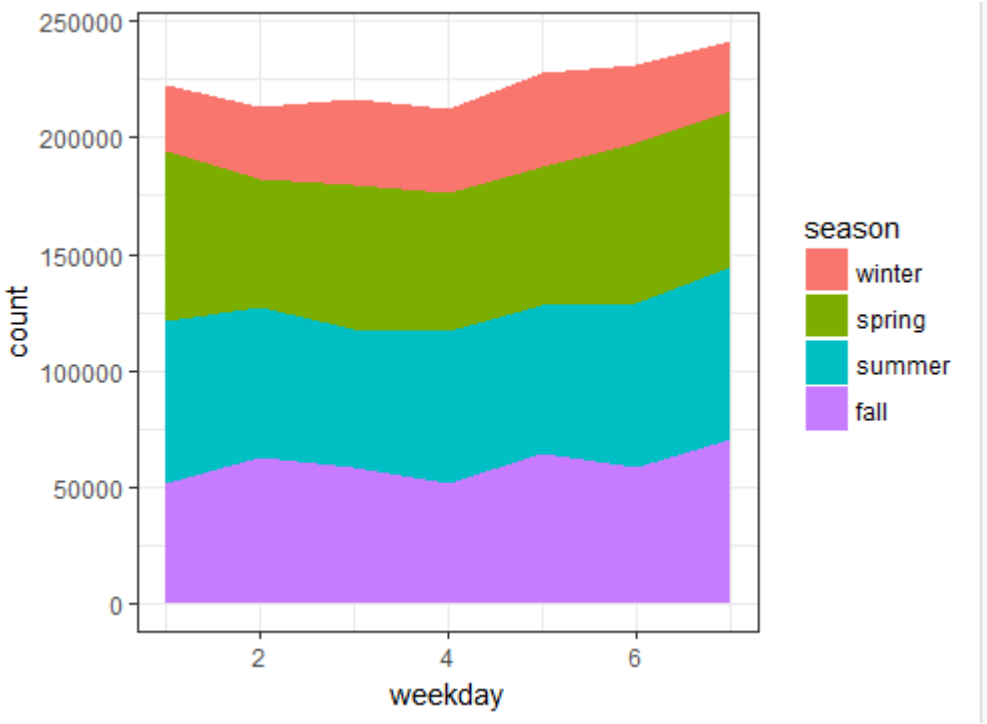
דרך נוספת להראות זאת הינה בגרף שטח **מבחינת העונות ביחס לשעות:**

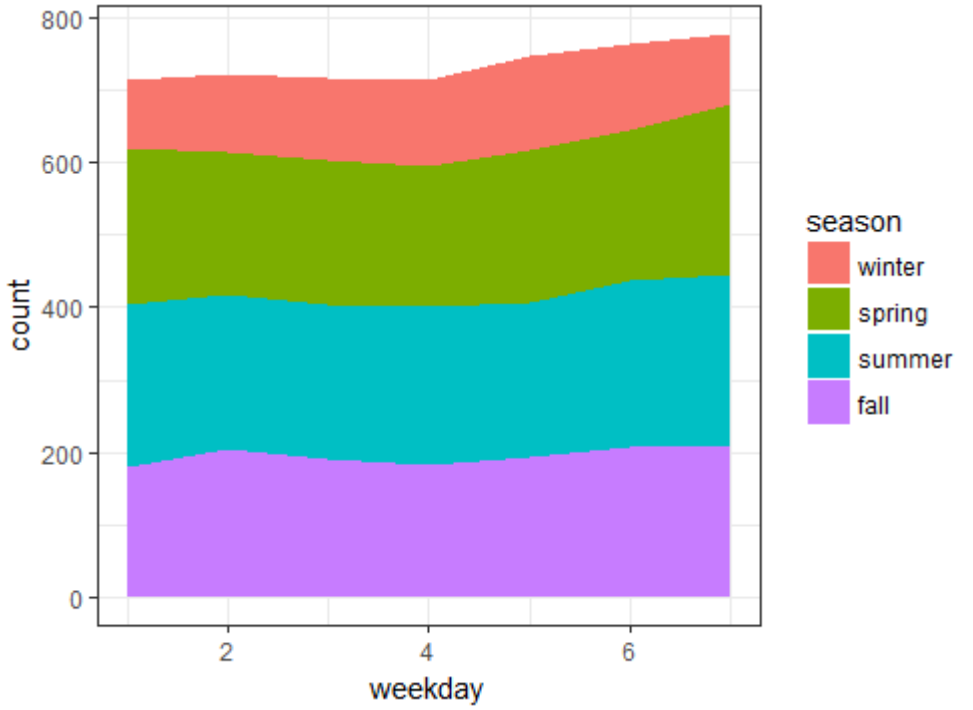




ניתן לראות **שאכן בקיץ ובאביב** ההשכרות הן הרבות ביותר במיוחד בשעות הנ״ל.

באופן דומה מבחינת **ימות השבוע** ולא שעות :

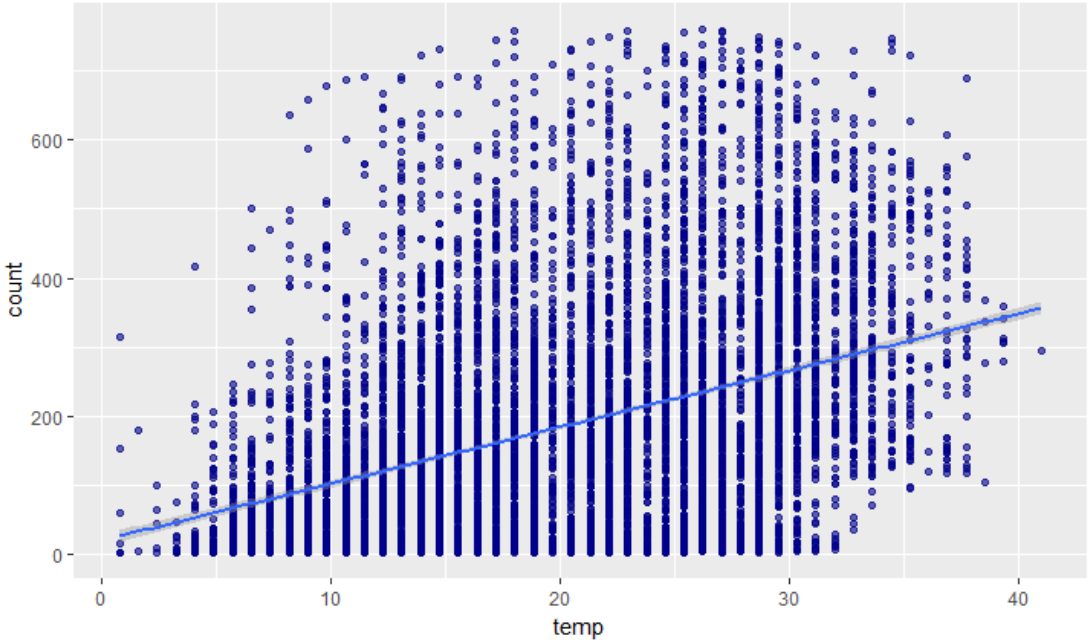




להראות מבחינת חודשים ועונות השנה אינו מוסיף מידע משמעותי.

2. <u>שאלה 2:</u>

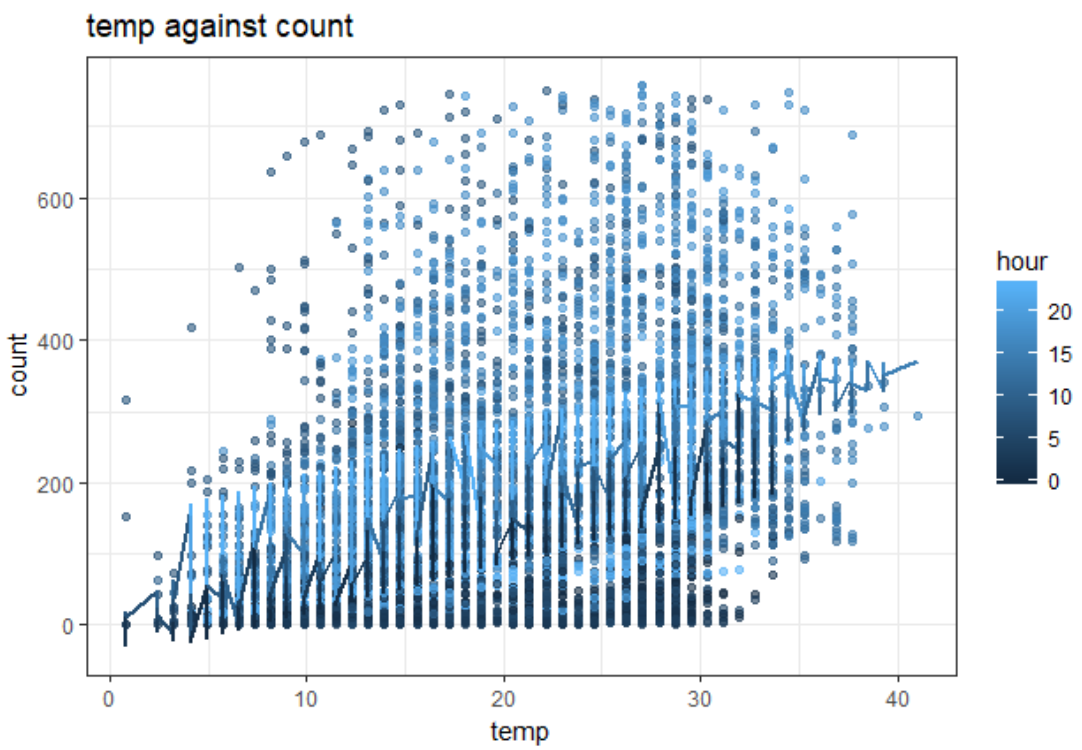
הרצנו רגרסיה לינארית - **כמות ההשכרות מוסברת על ידי טמפרטורה**:



חילקנו את התצפיות, 70% לtrain ו 30% ל validation.

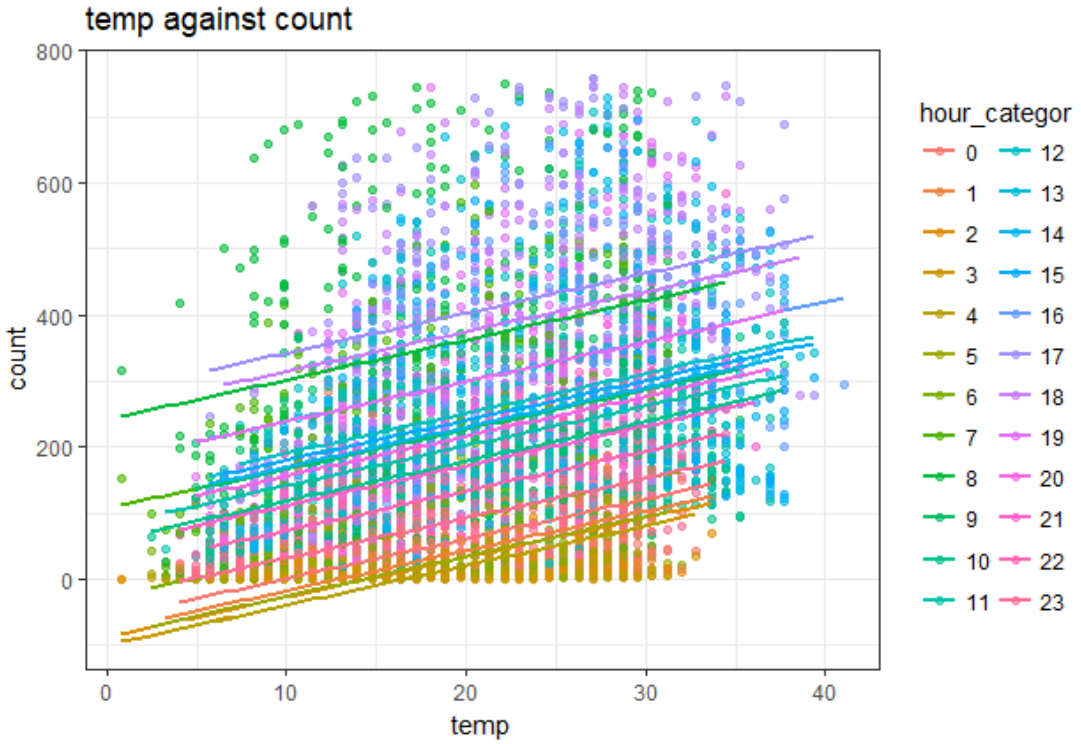
המודלים הבאים מבוססים על ה train set.

הרצנו רגרסיה לינארית רבת משתנים - **כמות ההשכרות מוסברת על ידי טמפרטורה וזמן ביום:**



משמעות המקדם של זמן ביום היא ההשפעה השולית של הזמן על כמות ההשכרות עבור טמפרטורה מסוימת (קבועה). משמעות זו **לא נשמעת לנו הגיונית** מפני שזמן ביום הוא לא משתנה שהערכים שלו מבטאים יחס של "גדול מ", כלומר זמן הוא חלוקה קטגוריאלית ואינו משתנה רציף.

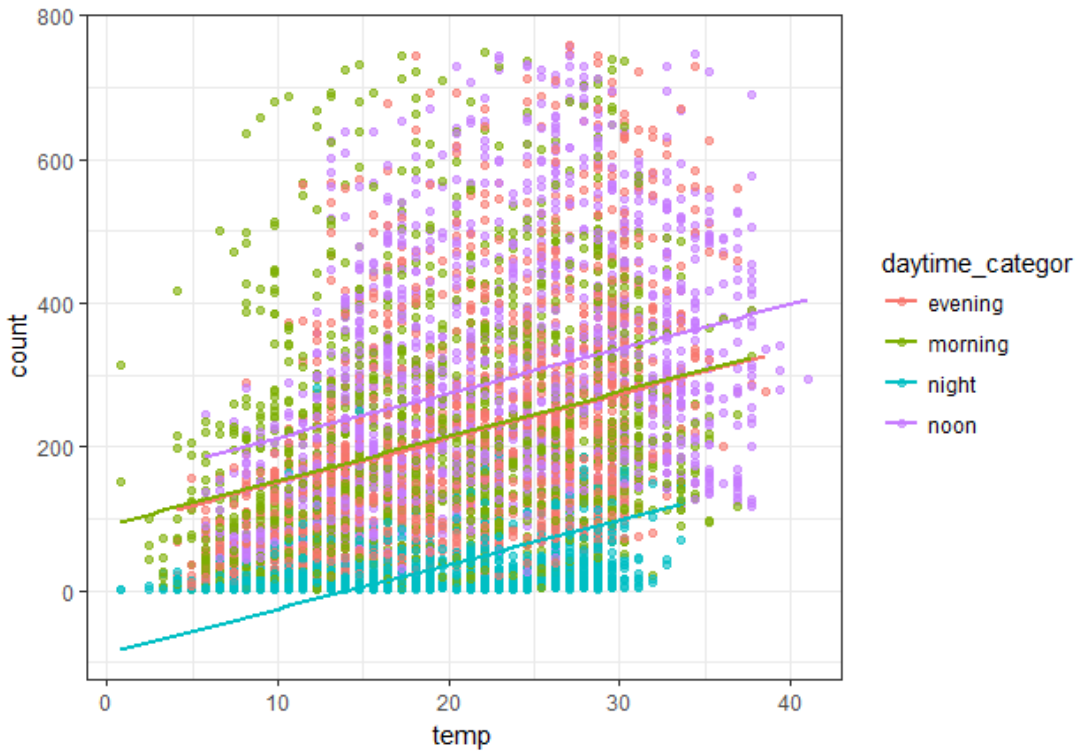
הרצנו רגרסיה לינארית רבת משתנים נוספת – **כמות ההשכרות מוסברת על ידי טמפרטורה <u>וזמן ביום כמשתנה קטגוריאלי</u>**:



כשמתייחסים לזמן ביום כמשתנה קטגוריאלי, הרגרסיה בעצם הופכת אותו ל-24 משתני דאמי, כלומר 24 רגרסיות ונותן זיהוי ייחודי לכל שעה.

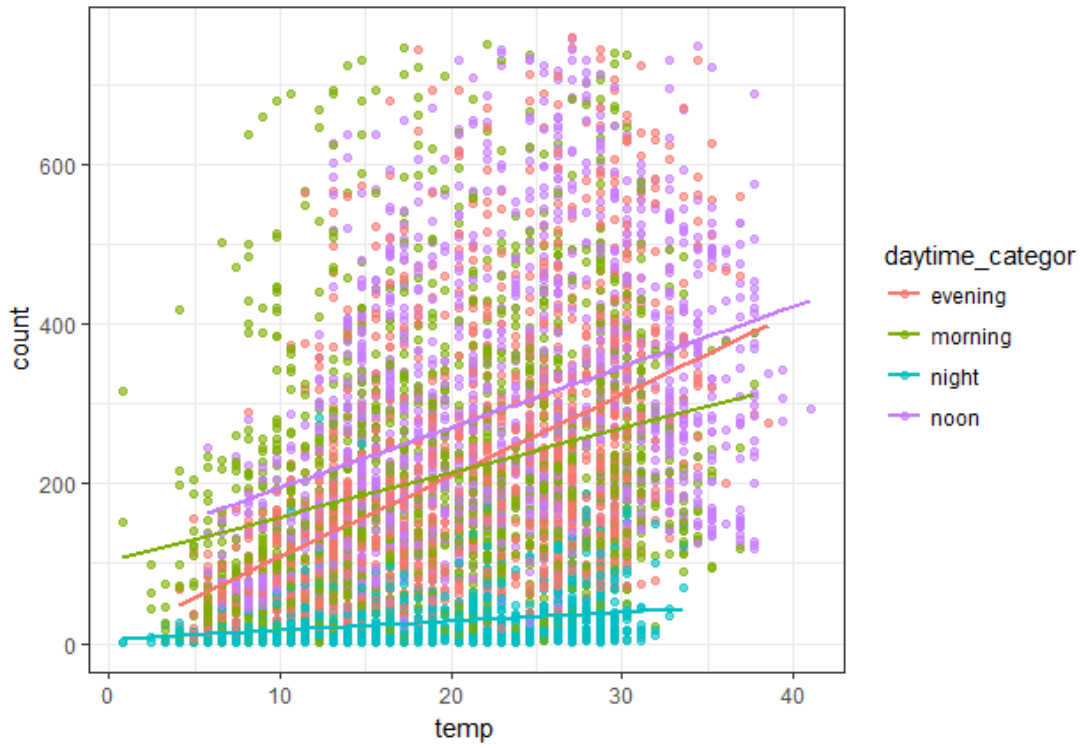
על מנת להביא לחיזוי טוב יותר, יצרנו משתנה קטגוריאלי חדש אשר אומר מה הוא החלק ביום. חילקנו את היום לארבעה חלקים – בוקר, צהריים, ערב ולילה. החלוקה התבצעה גם על סט של הtrain וגם על הסט של ה validation.

הרצנו רגרסיה לינארית רבת משתנים – **כמות ההשכרות מוסברת על ידי טמפרטורה <u>והחלק ביום</u>**:



נראה רגרסיה לינארית עם משתנה קטגוריאלי. כמו ברגרסיה הקודמת, גם כאן יש התנהגות של משתנה דאמי, אך במקום ריבוי רמות למשתנה זמן, איחדנו את הרמות לארבעה חלקים של זמן במהלך יממה.

הרצנו רגרסיה לינארית רבת משתנים נוספת – **כמות ההשכרות מוסברת על ידי <u>האינטראקציה</u> בין טמפרטורה והחלק ביום**:



במודל עם אינטראקציה אנחנו מאפשרים להשפעה של טמפרטורה להיות שונה לכל חלק ביום (נוכל לראות זאת בשיפועים השונים).

R^2 לארבעת המודלים שהרצנו :

```
> summary(multi_reg_daytime)$r.squared
[1] 0.4108927
> summary(multi_reg_daytime_interaction)$r.squared
[1] 0.432415
> summary(multi_reg)$r.squared
[1] 0.5858885
> summary(multi_reg_categorical)$r.squared
[1] 0.5757509
```

נראה ש R^2 למודל שמסביר כמות השכרות על ידי טמפרטורה וזמן ביום כמשתנה רציף, הוא הגבוה ביותר. המשמעות של הדבר הינו שמודל בעל ה-R^2 הגבוה ביותר הוא כנראה המודל שכנראה עשוי (לא בטוח!) להתאים בצורה הטובה ביותר ולחזות לנו באופן מדוייק יותר.

למרות זאת, המודל עם האינטראקציה שיש לו R^2 נמוך יותר, אינו בהכרח מודל גרוע יותר. באופן כללי הסיבה לכך היא שייתכנו מקרים בהם יהיה overfitting של המודל לנתוני ה-train ונקבל מודל עם R^2 גבוה יותר, אך כשנבדוק את הפרדיקציה על ה- test set נקבל תוצאות טובות יותר שתואמות למציאות.

```
> print(validation_sse_daytime_interact)
[1] 40354336
> print(validation_sse_daytime)
[1] 42569013
> print(validation_sse)
[1] 51170935
> print(validation_sse_categor)
[1] 29274258
```

במקרה שלנו, אחרי בדיקת ולידציה (בדיקת התאמת הפרדיקציה לטסט סט) קיבלנו שהמודל בעל R^2 הגבוה ביותר (כאשר המשתנה הוא רציף), הוא גם בעל ה SSE הגבוה ביותר ולכן אינו המודל הטוב ביותר.

המודל המסביר כמות השכרות לפי טמפרטורה וזמן ביום כמשתנה קטגוריאלי, המודל הטוב ביותר, בעל ה SSE הנמוך ביותר.

3. שאלה 3 :

:Model

1. עבור המשתמש agg_climate יצרנו את המודל הבא:

- ראשית, השתמשנו בנוסחא שמצאנו במחקר אודות Australian apparent temperature אשר נותן atemp כתוצאה מטרנספורמציה ושקלול של המשתנים relative humidity, temperature ו-windspeed.
להלן מתוארת דרך החישוב:

"

AT = Ta + 0.33×e − 0.70×ws − 4.00
Ta = Dry bulb temperature (°C)

e = Water vapour pressure (hPa) [humidity]

ws = Wind speed (m/s) at an elevation of 10 meters

The vapour pressure can be calculated from the temperature and relative humidity using the equation :

e = rh / 100 × 6.105 × exp (17.27 × Ta / (237.7 + Ta)) "

- כמו-כן פירקנו כל משתנה קטגוריאלי (weather ו-season) למשתני דאמי בינאריים של 0 ו-1. בסה"כ קיבלנו 8 משתנים עבור 2 משתנים קטגוריאלים.
- מכיוון שישנה קורלציה של 0.99232466 בין המשתנה atemp למשתנה temp ניתן להסיר אחד מהם. אולם, בחרנו לעשות ממוצע ל-atemp ולaustralian_atemp בכדי להגיע למשתנה יחיד וכי איננו יודעים כיצד חושב ה-atemp ב-data set. בנוסף, נרמלנו את הנתון הזה שיהיה בין 1-ל-1 ע"י השיטה:
(value -min(value)/(max(value)-min(value).
- לבסוף חיברנו את המשתנים הבינאריים יחד עם ה-normal_avg_temp להיות המשתנה agg_climate.
- מכיוון שהיה נראה ברגרסיה כי ישנה שונות שונה, הרצנו WLS במקום OLS באמצעות מתן משקל של weight=1/ agg_climate^2.
- הנרמול נותן בין 0 ל 1 וכל משתנה קטגוריאלי נותן 1 או 0. מפני שיש לנו 2 משתנים קטגוריאלים שכל אחד מהם נותן 0 או 1, אז כל רשומה יכולה לקבל בין 2 ל 3. כך קיבלנו משתנה מנורמל שנע בין 2ל-3:

```
summary(bike_train_clean_train$agg_climate)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.052   2.319   2.464   2.459   2.598   2.914
```

הרצנו רגרסיה:

```
Call:
lm(formula = count ~ agg_climate, data = bike_train_clean_train,
    weights = 1/(agg_climate^2))

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-102.51  -44.49  -13.50   32.52  260.30

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -761.43      29.18  -26.10   <2e-16 ***
agg_climate   384.96      11.95   32.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 62.87 on 5957 degrees of freedom
Multiple R-squared:  0.1484,    Adjusted R-squared:  0.1483
F-statistic:  1038 on 1 and 5957 DF,  p-value: < 2.2e-16
```

2. השתמשנו ב-hour categorial שזה משתנה השעות בצורה קטגוריאלית (אשר מפורש כמשתני דאמי בעת הרגרסיה).

```
Residual standard error: 111.5 on 5934 degrees of freedom
Multiple R-squared:  0.5678,    Adjusted R-squared:  0.5661
F-statistic: 324.9 on 24 and 5934 DF,  p-value: < 2.2e-16
```

והרצנו רגרסיה:

```
Call:
lm(formula = count ~ agg_climate + hour_categor, data = bike_train_clean_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-368.70  -62.83   -6.10   51.30  484.49 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -612.711     22.870 -26.791  < 2e-16 ***
agg_climate     274.798      8.957  30.679  < 2e-16 ***
hour_categor1   -18.512      9.845  -1.880  0.06011 .  
hour_categor2   -28.322      9.886  -2.865  0.00419 ** 
hour_categor3   -40.802      9.917  -4.114 3.93e-05 ***
hour_categor4   -41.620     10.017  -4.155 3.30e-05 ***
hour_categor5   -26.761     10.040  -2.665  0.00771 ** 
hour_categor6    29.950      9.902   3.025  0.00250 ** 
hour_categor7   164.847      9.857  16.724  < 2e-16 ***
hour_categor8   298.590      9.906  30.141  < 2e-16 ***
hour_categor9   163.962      9.827  16.685  < 2e-16 ***
hour_categor10  115.676      9.799  11.805  < 2e-16 ***
hour_categor11  139.294      9.739  14.302  < 2e-16 ***
hour_categor12  189.143      9.837  19.229  < 2e-16 ***
hour_categor13  178.088      9.993  17.821  < 2e-16 ***
hour_categor14  167.017      9.812  17.021  < 2e-16 ***
hour_categor15  177.881      9.955  17.868  < 2e-16 ***
hour_categor16  236.800      9.795  24.177  < 2e-16 ***
hour_categor17  339.411     10.217  33.219  < 2e-16 ***
hour_categor18  310.817     10.151  30.621  < 2e-16 ***
hour_categor19  235.906      9.894  23.844  < 2e-16 ***
hour_categor20  153.627      9.849  15.598  < 2e-16 ***
hour_categor21  109.298      9.770  11.188  < 2e-16 ***
hour_categor22   71.631      9.867   7.259 4.39e-13 ***
hour_categor23   30.848      9.875   3.124  0.00179 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 111.5 on 5934 degrees of freedom
Multiple R-squared:  0.5678,	Adjusted R-squared:  0.5661 
F-statistic: 324.9 on 24 and 5934 DF,  p-value: < 2.2e-16
```

חישוב ידני:

מכיוון שהחישוב ארוך (קיימים 24 משתנים) נראה את הדרך לחשב:

$$predictedResult_i = \beta_0 + \beta_1 * x_{1i} + .. + \beta_2 * x_{2i} + \beta_3 * x_{3i} + \cdots \beta_4 * x_{4i}$$

זאת כאשר agg_climate=b1
ו hour_categor1=b2 וכן הלאה.

לדוגמא, אם ב-Dataset המקורי ראינו את הנתונים הבאים:

Precidted value	agg_climate	hour
86.57244064	2.284051	**22**
45.80690659	2.284115	**23**

$$predictedResult_i = \beta_0 + \beta_1 * x_{1i} + \beta_{19} * x_{19i} = -612.711 + 274.798 \times 2.284051 + 71.631 = 86.572$$

$$predictedResult_i = \beta_0 + \beta_1 * x_{1i} + \beta_{20} * x_{20i} = -612.711 + 274.798 \times 2.284115 + 30.848 = 45.807$$

ניתן לראות כי התוצאות אכן דומות לערך החזוי.

- יש לשים לב כי תוצאות בערך החזוי שהחזירו ערך שלילי (count<0) – הפכנו אותם ל-0 כי לא ייתכן count שהינו שלילי.