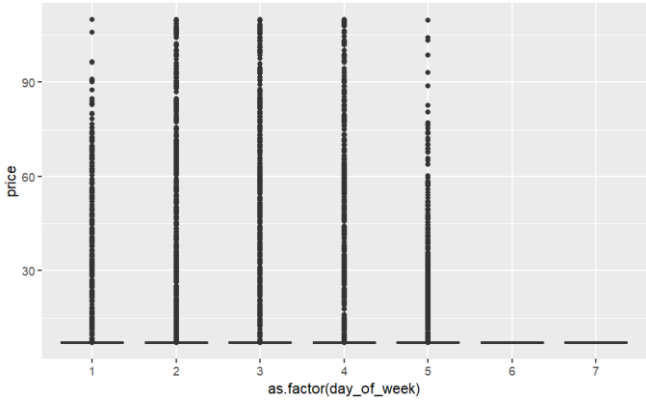
Yechiel Dan 308556364

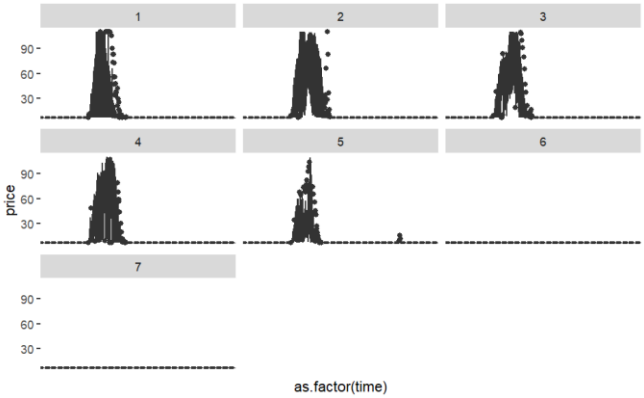
מטלה שלישית - חיזוי עתיד
כלכלה בעולם ה Big Data

חלק 1 - הכנת וניתוח הנתונים

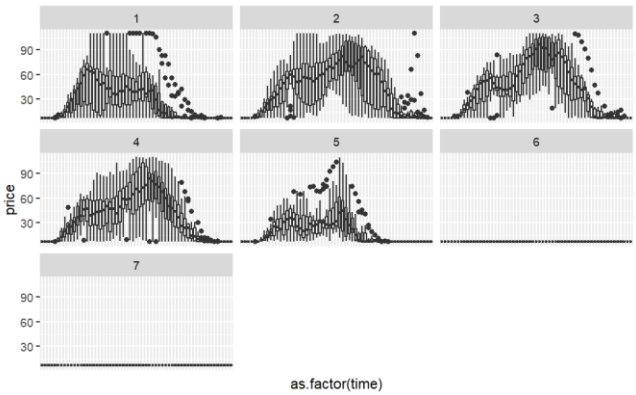
על מנת לחזות את עלויות הנסיעה בכביש המהיר בתאריכים 1/9/22 עד ל14/9/22, דבר ראשון ארצה לנתח ולהבין את סט הנתונים הקיימים לפני ביצוע אימון למודל וצפיית התחזית, לכן לאחר טעינת הנתונים המרכזיים, ניקויי שלהם והוספת נתוני מזג האוויר אציג את פלטי הנתונים בגרפים, על מנת להבין שינויי מאקרו ומיקרו על חודשים ימים שעות ודקות כאשר זהו סט הנתונים המרכזי והחשוב ביותר לצפיית העלויות הרצויות.



בגרף זה אנו רואים את חלוקת המחירים לפי ימים בשבוע, כאשר מה שאני מסיק מגרף זה שבסוף השבוע המחירים מינימליים בכל שעות היום, ובאמצע השבוע יש את המחירים הגבוהים ביותר – שזה מצביע על ביקוש יתר בכניסה לנתיב המהיר באמצע השבוע.



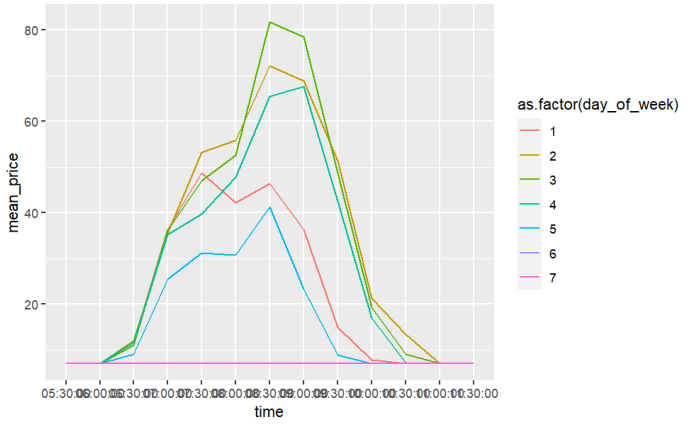
כאן אנו רואים את התפלגות המחירים כתלות בזמן מפורקים על פי ימי השבוע כדי להתחיל לבסס את ההשערה שהגורם העיקרי לעומס בנתיב המהיר הוא ההגעה לעבודה בתל אביב בשעות הבוקר (כאשר כאן אפשר לראות שבימי שישי שבת ואין עבודה, אין ביקוש לנסיעה בנתיב המהיר).



זהו אותו גרף כמו הקודם אך לאחר מחיקת התצפיות אשר לא בשעות העומס ב'בוקס פלוטים' – כדי לראות יותר מדוייק את השעות הרלוונטיות על פי ימים.

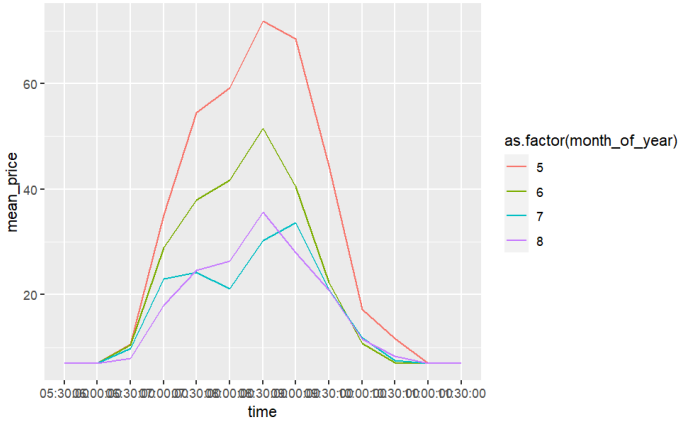
בגרף זה אפשר לראות את ימי השבוע בצבע נפרד לכל יום עם מחיר ממוצע לפי השעות.

כאשר מראה שבין השעות 8:00 ל9:00 יש את העומס המירבי ועל כן גם המחיר בנתיב הוא המירבי בכל יום.

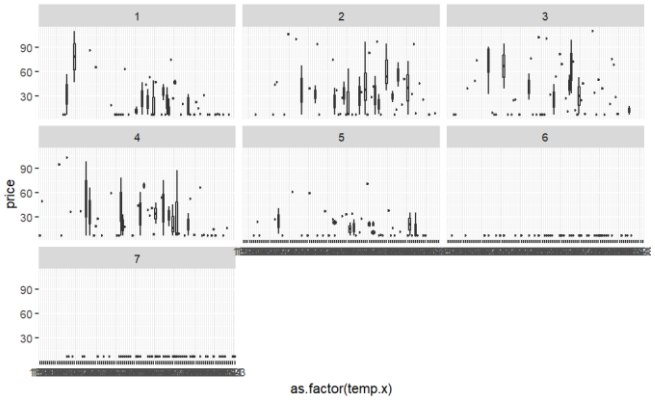


כאן אנו רואים את החודשים בנפרד, כך שלכל חודש המחיר הממוצע לפי השעות.

ומראה שככל שאנו מתקרבים לחודשי הקיץ ולחופש המחיר יורד – זאת אומרת פחות ביקוש לנסיעה בנתיב המהיר.



בגרף זה הצגתי את המחיר לפי הטמפרטורה אך מחולקת לפי ימי השבוע, כאשר רציתי לבדוק האם יש קשר מסויים שאוכל להיעזר בו, אך מכיוון שאני לא רואה איזה קו מגמה כלשהו בין הטמפרטורה זה גרם לי להוריד מהחשיבות של הטמפרטורה על המחיר.



עוד משתנה שלדעתי היה יכול להיות רלוונטי לתחזית, היה ימי חופש כללים/ חגים באמצע שבוע וכדו' כדי לתת תחזית מדוייקת יותר, אך בבדיקה על ימי התחזית אין בהם ימי חופש/ ימים שונים מבחינת עבודה, ולכן לא הכנסתי זאת.

לכן מכלל הנתונים אני מסיק שהגורם המרכזי למחיר הוא כמובן השעה, אך בינתן היום בשבוע שמשפיע ביותר, מכיוון שמהניתוח וההבנה שלי יש עומס גדול יותר בימי העבודה בשבוע ולפי זה גם בבקרים שבהם יש ביקוש רב לנתיב המהיר – שאלו בדיוק שעות העבודה בימי העבודה.

חלק 2 – הכלכלן

על פי הנתונים שנותחו ולדעתי המשתנה המרכזי שלפיו ארצה לבדוק את הקשר הסיבתי הוא יום בשבוע.

וכדי לבחון קשר זה ארצה לבנות מודל שבו אכניס את משתנה המטרה שלי כמחיר הנסיעה בנתיב, ומשתני האינטראקציה בין יום בשבוע עם השעה והשעה בריבוע. כאשר התוספת של השעה בריבוע נובעת מהסתכלות על צורת הגרפים (מחיר על פני זמן עבור כל יום) – שבאיזור השעה 6:00 עולה עד למקסימום באיזור 9:00 ויורד בחזרה ב11:00 למחיר ההתחלתי.

נראה כי משתני האינטראקציה בין השעה בריבוע ורוב הימים הם שונים ומובהקים (למעט שישי ושבת שהם זהים כי לפי מה שראינו כבר לפני על אותו מחיר כל היום) ולכן מראים שקיים קשר סיבתי בין הימים והשעות למחיר הנסיעה בכביש המהיר.

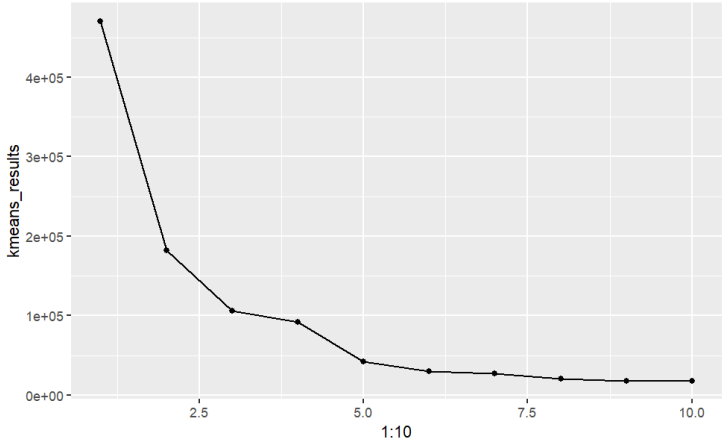
```
Call:
lm(formula = price ~ hour * as.factor(day_of_week) + hour_2 *
    as.factor(day_of_week), data = by_hour)

Residuals:
    Min      1Q  Median      3Q     Max
-14.272  -5.452  -0.097   0.000  95.559

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      7.69701    2.02728   3.797 0.000151 ***
hour                             1.43542    0.40990   3.502 0.000472 ***
as.factor(day_of_week)2         -0.68901    2.86244  -0.241 0.809806
as.factor(day_of_week)3         -1.59699    2.81797  -0.567 0.570971
as.factor(day_of_week)4         -0.59992    2.80906  -0.214 0.830907
as.factor(day_of_week)5         -0.27080    2.81997  -0.096 0.923508
as.factor(day_of_week)6         -0.69701    2.88061  -0.242 0.808830
as.factor(day_of_week)7         -0.69701    2.86606  -0.243 0.807878
hour_2                          -0.07406    0.01720  -4.305 1.75e-05 ***
hour:as.factor(day_of_week)2     1.27641    0.57636   2.215 0.026900 *
hour:as.factor(day_of_week)3     1.55676    0.57456   2.709 0.006796 **
hour:as.factor(day_of_week)4     0.75775    0.56740   1.335 0.181873
hour:as.factor(day_of_week)5    -0.52309    0.57252  -0.914 0.361008
hour:as.factor(day_of_week)6    -1.43542    0.58415  -2.457 0.014084 *
hour:as.factor(day_of_week)7    -1.43542    0.57916  -2.478 0.013277 *
as.factor(day_of_week)2:hour_2 -0.06165    0.02413  -2.555 0.010691 *
as.factor(day_of_week)3:hour_2 -0.07257    0.02411  -3.010 0.002647 **
as.factor(day_of_week)4:hour_2 -0.03652    0.02383  -1.533 0.125499
as.factor(day_of_week)5:hour_2  0.02687    0.02409   1.115 0.264849
as.factor(day_of_week)6:hour_2  0.07406    0.02452   3.021 0.002552 **
as.factor(day_of_week)7:hour_2  0.07406    0.02435   3.041 0.002385 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.42 on 2006 degrees of freedom
Multiple R-squared:  0.1508,    Adjusted R-squared:  0.1423
F-statistic: 17.81 on 20 and 2006 DF,  p-value: < 2.2e-16
```

```
        hour day_of_week     price
1  8.531915    2.893617 81.517554
2  8.059829    2.940171 36.856593
3  4.269677    4.201290  7.465058
4 17.074449    4.008272  7.008804
```



וכאן בניתוח קלאסטרינג אנו רואים בחלוקה ל4 קבוצות שהמחירים המקסימליים מתקבלים בבוקר ובתחילת השבוע.

חלק 3 - האם הנסיעה בנתיב המהיר משתלמת?

על מנת לחשב את כמות הזמן של נסיעה על הנתיב לחסוך לעומת אלטרנטיבת הנסיעה בנתיב רגיל, ארצה לשאול האם ההפרש במחיר כפול שכר לשעה גדול מהמחיר שעולה לנסוע בנתיב המהיר באותה שעה, בחלוקה של המחיר בשכר אמצא את הזמן בדקות שהנתיב המהיר צריך לחסוך למר ישראלי על מנת שיהיה משתלם לו לשלם עבור הנסיעה, עבור כל שעה עגולה מ7:00 עד 11:00 :

time	mean_price	time_be_saved_in_min
<chr>	<dbl>	<dbl>
07:00:00	20.16667	12.10000
08:00:00	41.36364	24.81818
09:00:00	35.00000	21.00000
10:00:00	10.50000	6.30000
11:00:00	7.00000	4.20000

חלק 4 – התחזית

על מנת לחזות בצורה המיטבית הממזערת את 'טעות ריבועית ממוצעת' הרצתי 2 מודלים על כאשר הראשון הוא ברגרסיה לינארית והשני הוא 'יער אקראי' עם המחיר כתלות באינטראקציה בין הזמן והיום בשבוע ובנוסף באינטראקציה של הזמן בריבוע והיום בשבוע.

המודל נבחר כך משום שבניתוח הנתונים לא ראיתי בהוספת הטמפרטורה מגמה ברורה שתעזור לי בניתוח, וגם במחשבה הגורם המרכזי לתחזית המחיר הוא היום בשבוע והשעה ביום ולכן התרכזתי בהם, כאשר לפי איך שראיתי את הנתונים הוספתי גם זמן בריבוע כי הגרף היה נראה בצורת פרבולה הפוכה.

```
lm_mod <- lm(price~time_as_int*as.factor(day_of_week) +time_as_int_2*as.factor(day_of_week), data =
train_df)
```

```
rndf <- randomForest(price~ day_of_week + time_as_int + time_as_int*as.factor(day_of_week),data =
train_df,mtry = 5,ntree = 1000)
```

אימון המודל היה בשיטת ה'קרוס ווליידציה', כאשר בדקתי את התחזית קודם על חלק מהנתונים שאותם לא הכנסתי לאימון המודל, ורק לאחר מכן על הנתונים שלא הוכנסו לאימון ביצעתי את הטסט ולאחר בדיקת טעויות הוצאתי את התחזית עבור חודש ספטמבר.

במודל 'יער אקראי' כי ה'טעות ריבועית ממוצעת' הייתה 256.0224 (גם לאחר שינוי כמות עצים ומשתנים), והטעות ברגרסיה הלינארית הייתה 363.9248 ולכן בסופו של דבר בחרתי במודל 'יער אקראי' שבה הטעות נמוכה יותר.