כלכלה בעולם הביג דאטא – מטלה 3

אורי שהם

מוחמד קיס

חלק 1 : סידור הנתונים וסטטיסטיקה תיאורית

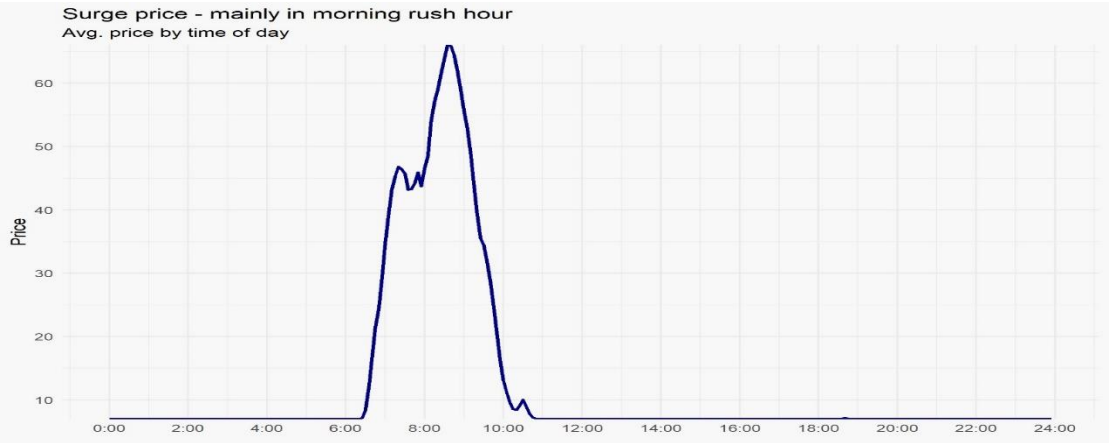
תחילה, הסתכלנו על הדאטא סט שלנו ושמנו לב לכמה דברים, מספר התצפיות לכל שעה הוא לא שווה לכל שעה לכן עשינו ממוצע המחיר כל 5 דקות כדי שנקבל אותו מספר תצפיות בכל שעה. דבר שני הוא שרוב התצפיות המחיר בהן הוא 7 לכן רצינו להסתכל על התפלגות המחיר כדי להבין מתי הוא עולה בדיוק והגדרנו משתנה surge price שהוא דמי מקבל 1 אם המחיר גדול מ7. ראינו גם שהמחיר בימי סופי השבוע וחגים הוא קבוע על 7, לכן הורדנו את הימים אלו מהדאטא שלנו כדי להבין יותר טוב מה משפיע על עליית המחיר[1].

משתנים מסבירים שהוספנו :

1- Surge_price : משתנה דמי מקבל 1 המחיר הוא מעל 7
2- Max_price : משתנה דמי מקבל 1 אם המחיר הוא 110
3- school_sabbatical : משתנה דמי מקבל 1 אם בתי הספר היסודיים היו בחופשה
4- לחות וטמפרטורה (נתונים 10 דקתיים מהשירות המטאורולוגי
5- משתנים מבוססי זמן (יום, שעה..)

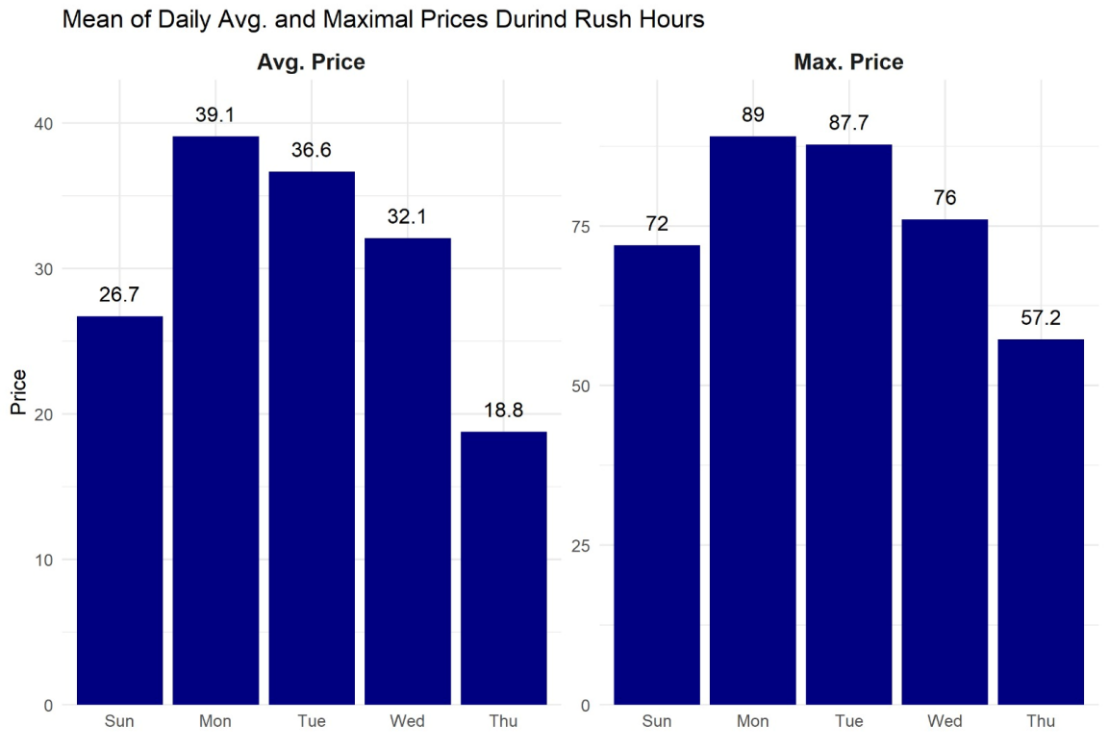
אנו מציגים כאן גרפים נבחרים מתוך הניתוח, הקוד המצורף כולל גרפים רבים נוספים בהם סטטיסטיקה תיאורית במשתנה יחיד, ניתוח נתוני מזה האוויר שלא הובילו לתוצאות מעניינות ובחינה של ימים חריגים שלא הובילה להחלטה לשנות את הנתונים. ניתן לראות גרפים אלו כאן.

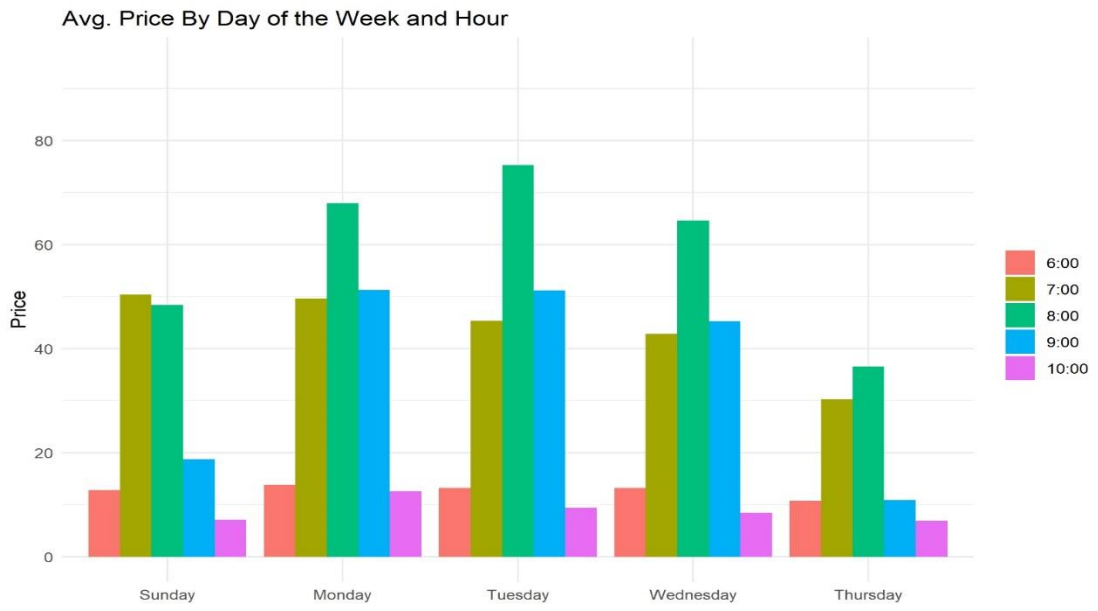
בגרף להלן רואים ששעות העומס או השעות שבהן המחיר עולה על 7 הן בין 6:00 עד 11:00 לכן החלטנו להוריד את השעות האחרות וצמצמנו את הדאטא לשעות 6:00 ל11:00 בלבד.



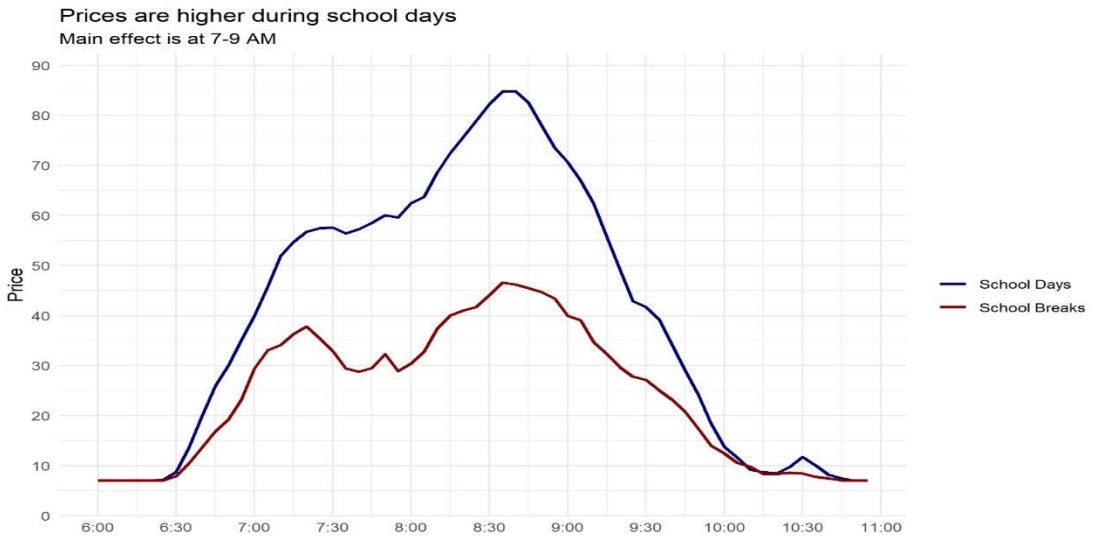
[1] גרף מופיע כאן.

בגרף להלן רואים את המחיר המקסימלי (הממוצע) לפי יום לעומת המחיר הממוצע (הממוצע) לפי יום, יכולים לראות שבימי חמישי המחירים הם הכי נמוכים לעומת ימי שלישי שהמחירים בהם הם הכי גבוהים.





בגרף למעלה רואים את המחיר הממוצע לפי יום בשעות הנתונים, רואים ששעות הפיק הן 7,8 ו 9 ולעומת שכל הימים נחשבים כימי אמצע שבוע יכולים לראות שבימי שני שלישי ורביעי המחירים הרבה יותר גבוהים מהמחיר בימי ראשון וחמישי. ניתן לראות גם שמסלול המחיר הממוצע הוא שונה בין הימים.



בגרף למעלה רואים את האפקט של ימי החופש של בתי הספר, בימי בית ספר המחירים מגיעים למספרים יותר גדולים בשעות הבוקר (שעות בית הספר) לעומת ימים בחופש והמחירים מתקרבים מחדש אחרי סיום שעות העומס. באופן כללי ניתן לראות שיש השפעה לימי הלימודים וגרפים נוספים משחזרים את הניתוח שעשינו קודם בחלוקה לימי לימודים וימי חופש וניתן לראותם כאן.

חלק 2: הכלכלן

ראינו שהעומס הנוצר בימי לימודים מתואם חיובית עם מחיר הנסיעה בנתיב המהיר וההינו רוצים לבחון את ההשפעה הסיבתית של בתי הספר בהקשר זה. ברור כי השפעת בתי הספר מוגבלת בשעות סביב תחילת הלימודים ולכן אם היה למו משתנה שמציין את שעות הנסיעות ללימודים היה ניתן להשתמש בו במולד הפרש ההפרשים בצורה הבאה:

$$Price_{dt} = \beta_0 + \beta_1 schoolDay_d + \beta_2 schoolHour_t + \beta_3 schoolDay_d * schoolHour_t + \gamma^T controls_{dt} + \epsilon_{dt}$$

כאשר:

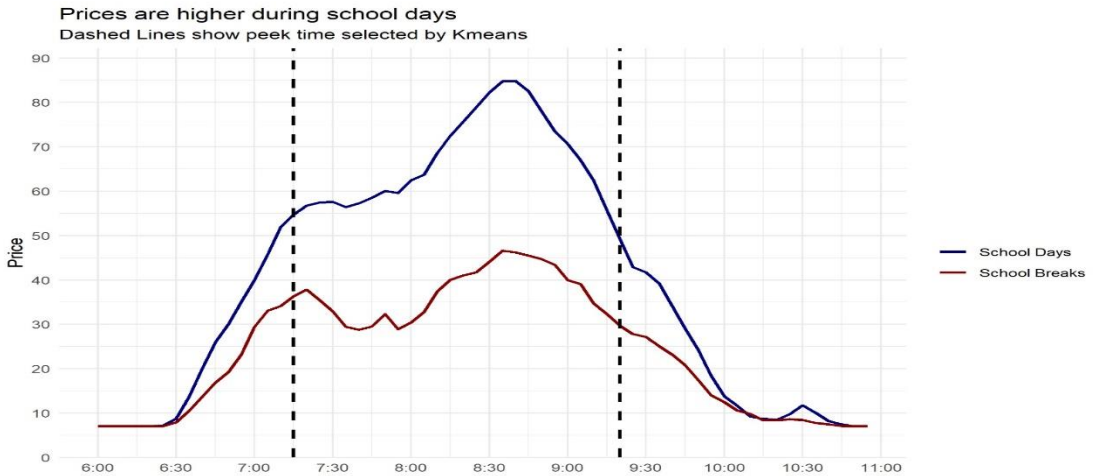
1- schoolDay: משתנה דמי מקבל אחד אם יום d הוא יום לימודים
2- schoolHour: משתנה דמי מקבל אחד אם שעה t היא שעת נסיעה לבית ספר
3- controls: וקטור של משתני בקרה (יכולים למשל לכלול יום בשבוע, מזג האוויר וכו')

אמידת β_3 מהמודל הייתה נותנת את ההשפעה הסיבתית של הנסיעות לבתי הספר (אפשר לחשוב על משתנה SchoolHour כמשתנה after במודל הפרש ההפרשים קלאסי ועל ימי בית הספר כקבוצה שעוברת את הטיפול).

מכיוון שאין לנו דרך טריוויאלית להגדיר את המשתנה SchoolHour נסינו לעשות זאת בעזרת אלגוריתם K – Means, הירצנו את האלגוריתם על המשתנים School_sabbatical ו price (לאחר סטנדרטיזציה של הנתונים). ראינו לפי כלל המרפק שמספר הקלסטרים המתאים הוא 4. ולאחר שהחזרנו את הנתונים לסקאלה המקורית אלו המרכזים שהתקבלו לקלאסטרים :

price	School_sabbatical
75.41	0
57.68	1
15.49	0
12.97	1

חיברנו את הקלסטרים לנתונים שכוללים גם א השעות ובחנו את פיזור השעות בכל קלאסטר, לאחר בחינה של הקלאסטרים ראינו שקלאסטרים 1 ו 2 תופסים את שעות העומס בימי לימודים ובימי חופש בהתאמה וקלאסטרים 3 ו 4 תופסים את השעות שלפני ואחרי שעות העומס. טווחי הזמנים שבהם נופלים 80% מהתצפיות של 1ו 2 מאוד דומים ולכן בחרנו להגדיר שעות בית הספר כטווח שנופלים בו 80% מהתצפיות של 1 ו 2 במשותף. הגרף שלהלן מראה את המחירים בחופש ובימי בית הספר והן את הטווח שנבחר בעזרת האלגוריתם והתהליך שתיארנו והוא נראה תואם להנחות מודל הפרש ההפרשים.



חלק 3 : האנליסט העסקי

מכיוון שאין בידינו נתונים רלוונטיים להערכת הזמן שייקח למר ישראלי להגיע מביתו לכניסה לנתיב המהיר נניח שהוא נכנס לנתיב המהיר בשעה עגולה t. הנסיעה בנתיב המהיר תשתלם למר ישראלי אם מתקיים :

$$S(t) * \left(\frac{100}{60}\right) \geq E\big(p(t)\big)$$

כאשר :

1 : S(t) – הזמן בדקות שנחסך בשעה t

2 : p(t) – המחיר בשקלים בשעה t (אומדים את התוחלת באמצעות הממוצע)

3 : 100/60 – השכר שלו בדקה בשקלים

שעה	הזמן המינימלי שהוא חייב לחסוך כדי שישתלם לו לנסוע הנתיב המיהר (בדקות)
7:00	23.4
8:00	33.6
9:00	37.6
10:00	7.8
11:00	4.2

שמנו לב שיש שונות ניכרת במחירים בין הימים השונים, לכן חילקנו את הניתוח גם לפי ימים והטבלה להלן מציגה את זמן החיסכון המינימלי בכל שעה עגולה בכל יום בשבוע :

	7:00	8:00	9:00	10:00	11:00
ראשון	26.6	33.1	27.7	4.6	4.2
שני	23.2	36.7	43.4	10.8	4.2
שלישי	23.8	41.3	52.1	10.2	4.2
רביעי	25.0	34.8	48.8	8.5	4.2
חמישי	18.5	21.2	14.2	4.2	4.2

חלק 4 : מדען הנתונים - פרדיקציה

ראשית חילקנו את הנתונים שלנו ל training set ול-validation set שכלל כ-20% מכלל התאריכים שקיימים בדאטה שלנו[2]. משתני מזג האוויר ב-validation set הוגדרו באמצעות ממוצע בארבע השנים הקודמות בכל זמן ביום ולא באמצעות נתוני האמת מהשנה. את המידע שהוקצה ל-training עיבדנו בשלוש דרכים שונות – ראשית (ungroup בקוד) מידע גולמי בו כל שורה מייצגת תצפית בודדת והמשתנים surge ו-max_price מייצגים מחיר גבוה מ-7 ₪ ומחיר של 110 ₪ בתצפית זו בהתאמה. בשתי הדרכים האחרות כל שורה מייצגת את המחיר הממוצע בטווח של חמש דקות כאשר בראשונה (any בקוד) המשתנים surge ו-max_price מקבלים ערך 1 אם בתצפית כלשהי

[2] בחרנו להקצות ל-validation set ימים שלמים מכיוון שמשימת התחזית היא לימים שלמים שאינם בדאטה ואחרת היה עשוי להיווצר חשש לoverfitting.

בחמש הדקות האלה המחיר היה גבוה מ-7 ₪ או 110 ₪ ובשנייה (all בקוד) המשתנים קיבלו 1 אם מחיר surge או מחיר מקסימלי נמשכו לכל אורך חמש הדקות.

בחנו שתי דרכים שונות לחזות את מחיר הנתיב המהיר:

1. מודל לינארי – לפי רשימת משתנים מסבירים אנו מתאימים מודל לינארי על ה training set- ומשתמשים בו לחזות את המחיר ב-validation set (test set בהמשך). אנו מתקנים כל מחיר שנחזה מתחתת 7 ₪ או מעל ל-110 ₪ למחיר הקצה הקרוב.
2. אמידה בשלבים – לפי רשימת משתנים מסבירים נתונה מבצעים ראשית קלסיפיקציה למחיר surge בעזרת מודל לוגיסטי, שנית קלסיפיקציה למחיר מקסימלי ולבסוף באמצעות מודל לינארי חוזים את המחירים לזמנים שלא סווגו כמחירי קצה. ראשית מבצעים קלסיפיקציה באמצעות מודל לוגיסטי למחיר surge. אנו עושים זאת באמצעות 5-fold cross validation על ה-training set לקביעת ערך הסף שימקסם את ה-accuracy ולאחר מכן התאמת מודל לוגיסטי על כלל ה-training set. תצפיות ב-validation set (test set בהמשך) עבורן נחזה מחיר מינימום נשארות בצד. באופן דומה מבצעים את הקלסיפיקציה למחיר המקסימום. לבסוף כדי לחזות מחירים לתצפיות שלא סיווגנו כמחירי קצה מורץ המודל הלינארי עם תיקון של תחזיות שחורגות מהתחום. שיטה זו משתמשת באותם משתנים מסבירים לכל השלבים מטעמי פשטות.

ייצרנו 44 קומבינציות של משתנים מסבירים שמבוססים על חופשות בתי הספר, יום בשבוע, הזמן ביום ומשתני מזג האוויר וכן פולינומים של הזמן ואינטראקציות[3]. משתנה הזמן הוגדר כמספר דקות מאז שש בבוקר (על משתנה זה גם הגדרנו פולינומים עד מעלה 3) או כמשתנה פקטור שמקבץ זמנים ביום לקבוצות של 5/10/15/20/60 דקות.

סך הכל היו לנו $2 * 3 * 44 = 264$ קומבינציות של דאטה לאימון, שיטת פרדיקציה ורשימת משתנים מסבירים[4]. בחנו את כל המודלים הללו וחישבנו עבורם את שורש ה-MSE (שהוא טרנספורמציה מונוטונית של ה-MSE) הן ב-training set והן ב-validation set. בחרנו את הקומבינציה שממזערת את ה-RMSE ב-validation set[5] ואנו חושבים שמודל זה אינו סובל מ-overfitting שכן ה-RMSE שלו דומה מאוד ב-validation set וב-training set[6]. את מודל זה אימנו מחדש בשיטת השלבים על כל הדאטה כאשר משתני surge ו-max_price הוגדרו לפי all והשתמשנו בו לחיזוי בתאריכי היעד. חזרנו על תהליך זה עם המודל הטוב ביותר שאינו משתמש במשתנה school_sabbatical למקרה שטרם ההגשה לא יהיה ידוע עדיין אם הלימודים יתחדשו בשל שביתות ולאפשר לבחור את התחזיות ברגע האחרון בהתאם להתפתחויות החדשותיות. כאן מופיעים גרפים דיאגנוסטיים על שגיאות המודלים שנבחרו והפרדיקציות שלהם ב-validation set וכן רשימת התוצאות של כל המודלים שנאמדו.

---

[3] בתהליך גיבוש רשימות המשתנים בחנו חלק מהמודלים מבלי להשתמש ב-validation set.

[4] למעשה מדובר בפחות שכן אין הבדל בין שני סוגי המידע המקובץ במודל הלינארי שאינו משתמש בסיווג של מחירי קצה.

[5] school_sabbatical*weekday_fac + min_10_fac*school_sabbatical , קיבוץ המשתנים והגדרת משתני דאמי לפי all ואמידה בשלבים.

[6] קיים חשש שמכיוון שנבחרו לנו בעיקר ימי חופש ל-validation set ובכלל מכיוון שספטמבר עשוי להיות שונה מהחודשים לפניו שכן מדובר ב-overfit מסוים לכלל התקופה. יתרה מזאת, המודל לא צפוי לדעת להתמודד עם ההשפעות של היום הראשון ללימודים שעשוי להיות יום חריג.