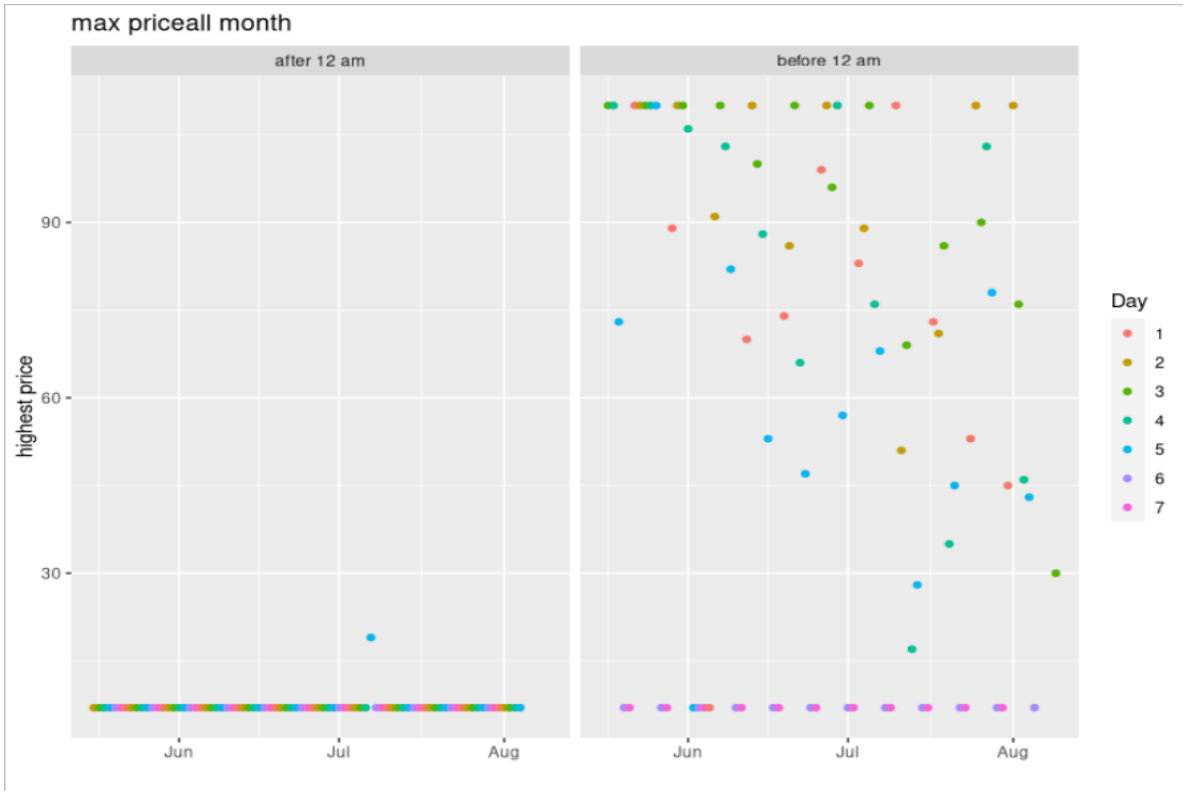
עמית שרון
דב טוך

כלכלה בעולם ה-Big Data – מטלה סופית

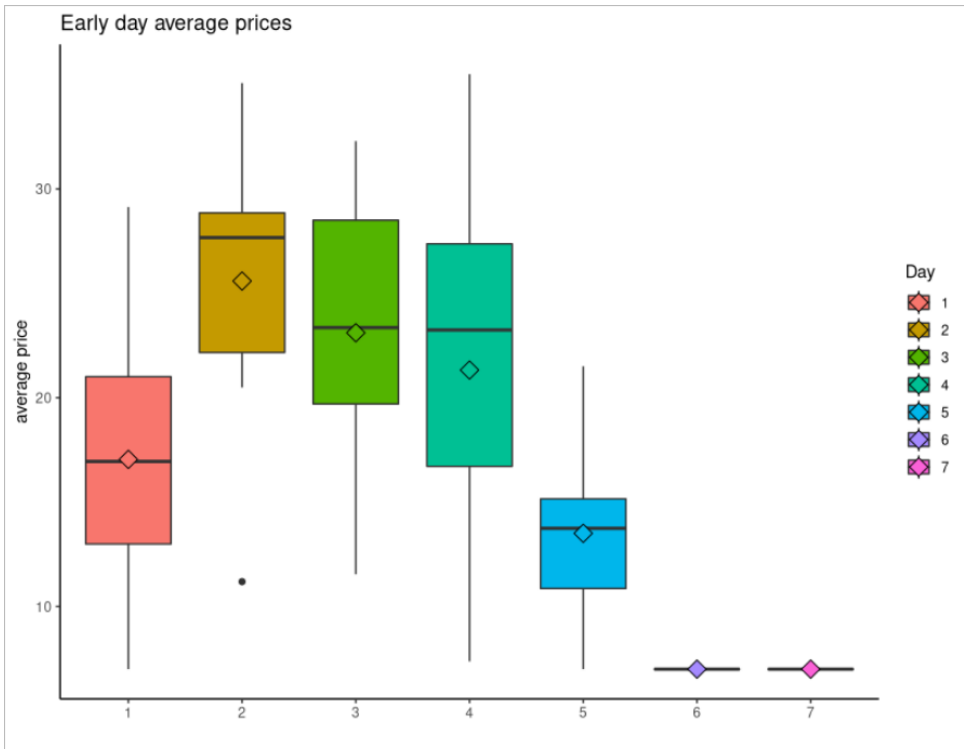
סידור וניקוי הנתונים: מפני שהנתונים מתעדכנים לכל הפחות בהפרשים של דקה החלתנו לשמור את הנתונים בהפרש של דקה ולא בהפרשים של) בערך(10שניות .יש לציין שמבקרה שבו היה שינוי בזמן בתוך שניות בתוך טווח הדקה (למשל אחרי השנייה ה30 חל שינוי במחיר הנסיעה ניקח את המחיר להיות זה שמייצג את הדקה הנ"ל.

3. החלטנו להוסיף את המשתנים: שעה, יום בשבוע, אינדיקטור לאירוע מיוחד באותו יום (חג, סגירה של רכבות/כבישים, ביקור הנשיא ביידן, פסטיבלים וכו'. אינדיקטור לכך שהשעה היא אחרי 12 בצהריים.

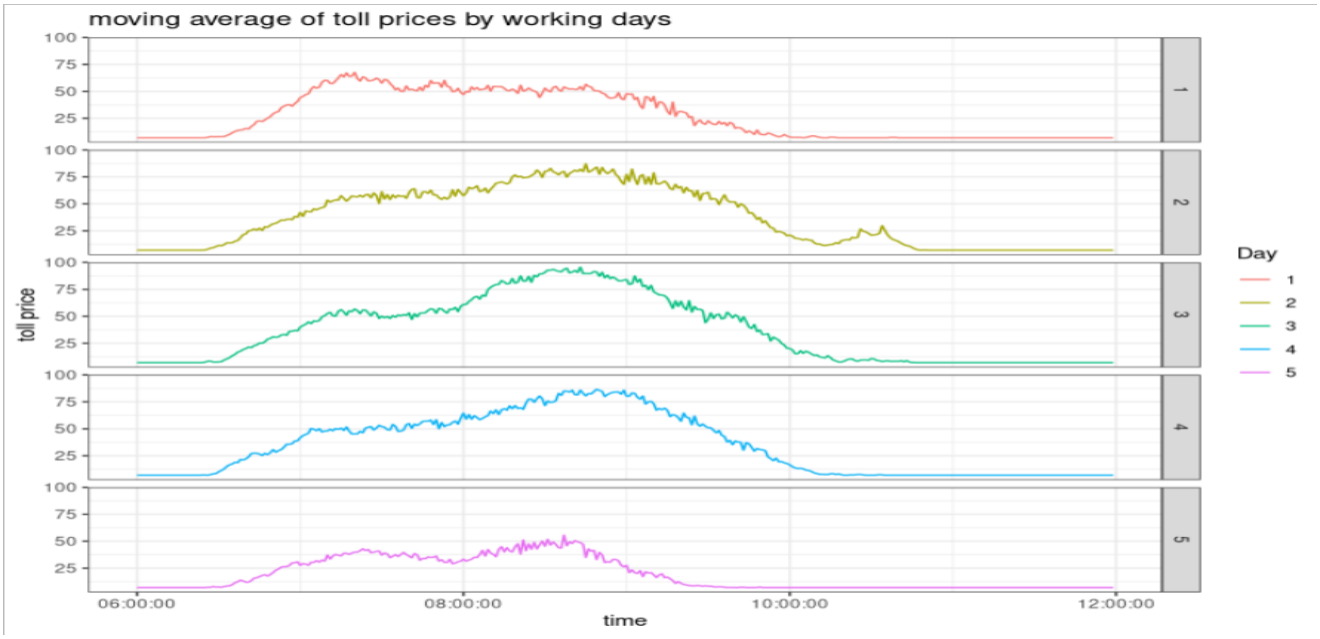


בגרף 1.1 ניתן לראות כי מחיר הנתיב משתנה כמעט לחלוטין בשעות המוקדמות של היום. לאחר השעה 12 פרט ליום אחד (שלא הצלחנו למצוא סיבה עבורו בחדשות) שבו המחיר המקסימלי בנתיב היה מעל מחיר המינימום והגיע ל-19 שקלים לנסיעה, אין כל שינוי במחירי הנטיב. על כן נמשיך לחקור את מחירי הנתיב בשעות המוקדמות של היום ובפרט בין השעה 6:00 ל-12:00.

עמית שרון
דב טוך



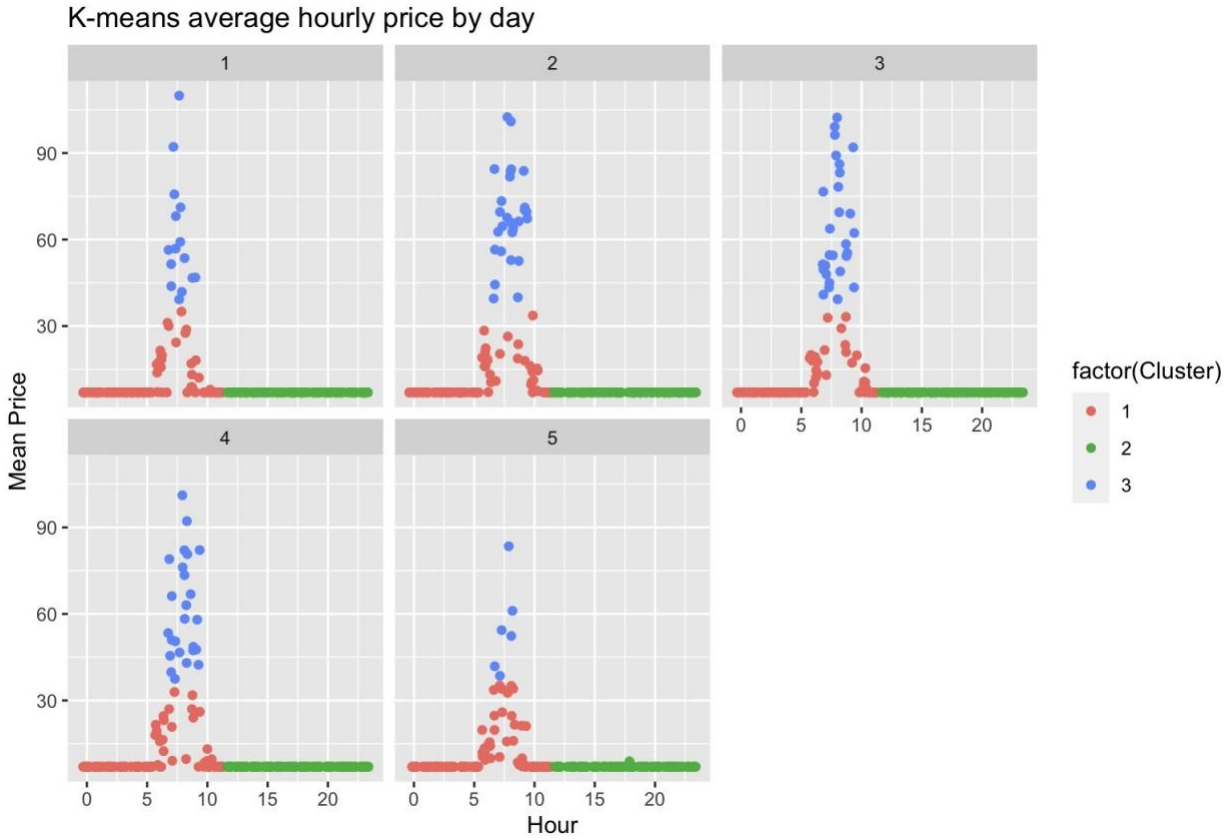
בגרף 1.2 לקחנו את המחיר הממוצע בשעות הבוקר של כל תאריך וסיווגנו אותם לפי ימות השבוע. שיערנו כי ימי ראשון וחמישי יהיו הימים הכי עמוסים אך הופתענו לגלות כי יום שני נמצא ביום שבו מחירי הנתיב המהיר הכי גבוהים. כמו כן בימי שישי ושבת מחיר הנתיב המהיר לא עולה מעל המחיר המינימלי (7) בכלל התצפיות שראינו (מסקנה שנשתמש בה בחלק ד׳ של החיזוי.) הסימון מסמל את ממוצע הממוצעים של התצפיות באותו היום.



בגרף 1.3 אנו מודדים את המחיר הממוצע בשעות 6:00 עד 12:00 בימים א-ה׳. ניתן לראות כי החל מהשעה

עמית שרון
דב טוך

6:30 לערך ישנה עליה במחיר הנסיעה. ביום א' העלייה הינה המהירה ביותר כאשר כבר לאחר שעה אנו מגיעים לשיא של כמעט 75 ₪ לנסיעה. בימים ב'-ד' ההגעה למחיר של 75 איטית יותר אך עוברת את ה75 לנסיעה כאשר השיא מגיע ב8:30 לערך. גם ביום ה' שיא המחירים מושג בשעה 8:40. בכלל הימים אחרי השעה 11:00 מחיר הנסיעה הינו תמיד המחיר המינימלי.



בשימוש במודל k means ניתן לראות כי המחיר הממוצע של התעבורה מתחלק ל-3 קלסטרים. מחיר בשעות הבוקר, בשעות שיא ומחיר ערב.

כפרשנות לאלגוריתם ניתן להבין שהשינוי במחירי הנתיב המהיר מושפע מהשעה ובמיוחד תלוי בשעות הבוקר ובסוג היום, סופש או יום חול, (בסופש המחיר אחיד כמעט, מפני שציינו זאת מקודם השמתנו זאת בגרף הנוכחי. כלומר יש קשר בין השעה ביום בשבוע למחיר ובעזרת שיטת IV נוכל גם למצוא גם קשר סיבתי לכך.

בחרנו במשתנה המסביר שעה ביום בשבוע. אפשר לגלות מה הקשר הסיבתי בינו לבין המחיר בנתיב המהיר ע"י שימוש instrumentale variable ו-sls2. משתנה שמתואם עם שעה ביום אבל לא עם מחיר הנתיב המהיר למשל מספר המוקדים שנותנים שירות פרונטלי בשעות מסויימות בלבד. הרעיון הוא ששירותי המדינה ובריאות פעילים בעיקר בין א-ה בין 8-13:30 ולכן יש מתאם עם יום ושעה אבל לא עם מחיר בנתיב.

עמית שרון
דב טוך

האנליסט העסקי:

חישבנו ב-R את העלות הממוצעת של נסיעה בנתיב בשעות בהן ישראל יכול לצאת מהבית:

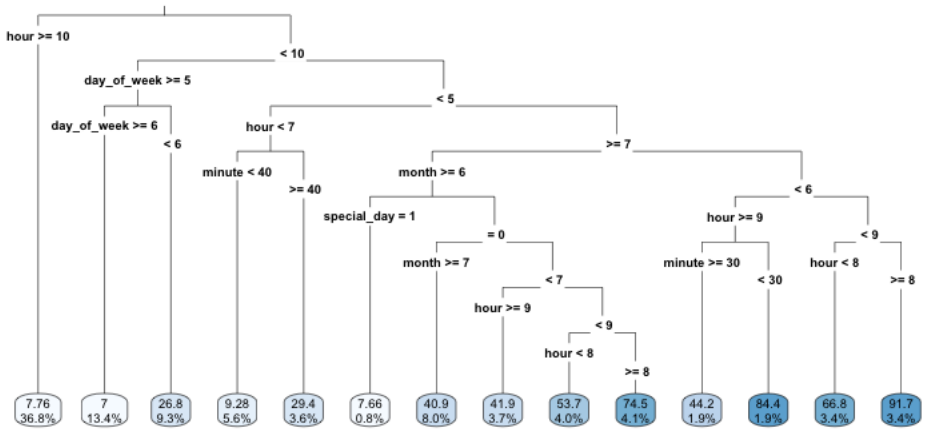
שעה	מחיר ממוצע באמצע השבוע
07:00	51.31 ש״ח
08:00	68.00 ש״ח
09:00	43.15 ש״ח
10:00	10.10 ש״ח
11:00	7.00 ש״ח

ישראל מרוויח 100 ש״ח לשעה, כלומר 1.67 ש״ח לדקה. כלומר, אם נסיעה בנתיב עולה x ש״ח, אז כדי שישתלם לישראל להשתמש בנתיב על הנסיעה בנתיב לחסוך לו $\frac{x}{1.67}$ ש״ח. לכן, נקבל את התוצאות הבאות כתלות בשעת היציאה של ישראל מהבית:

שעה	דקות נסיעה שהנתיב צריך לחסוך לישראל על מנת שישראל ישתמש בו
07:00	31
08:00	41
09:00	26
10:00	6
11:00	5

עמית שרון
דב טוך

<u>מדען הנתונים :</u>



ניסינו לחזות את המחירים באמצעות רגרסיה לינארית ובאמצעות עץ החלטה. קיבלנו תוצאות טובות יותר בעץ ההחלטה ולכן החלטנו לבצע את החיזוי באמצעות מודל זה.

ביצענו K-fold cross validation עבור מודל עץ ההחלטה כדי לוודא שאין overfitting וקיבלנו RMSE שדומה למודל לפני ביצוע הולידציה מה שמראה שגם עבור חלקים אחרים של הדאטה קיבלנו את אותה שגיאה ושהמודל חוזה בצורה טובה את המחיר.