הפקולטה להנדסה
ע"ש איבי ואלדר פליישמן
אוניברסיטת תל אביב

קורס שיטות אנליטיות בתמחור דינמי
Elements of Pricing Analytics

עבודה מסכמת:

נתוני מכירות ומחירי האבוקדו בשוק בארה"ב

מגישים:

אפק אדלר ת.ז: 204249239
קובי אדרי ת.ז: 021981147
כפיר סיטון ת.ז: 305007817
הדר שטרן ת.ז: 302883509
גלעד בוסי ת.ז: 031691082

מרצה:

פרופ' אסף זאבי

תוכן עניינים

רקע

שוק האבוקדו בארה"ב - Highlights

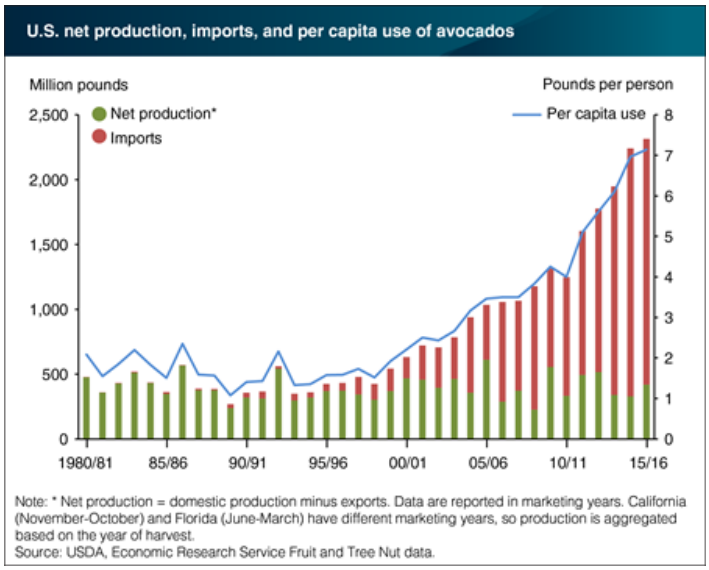
על מנת להבין את הנתונים ואת המגמות בשוק זה, נציין בקצרה מספר נקודות אשר יכולות להניח את הדעת על חלק מהדברים שגילינו בהמשך:

1) בארצות הברית ישנם שלושה אזורי גידול מרכזיים של אבוקדו: פלורידה, קליפורניה והוואי.

2) קליפורניה לבדה מייצרת כ 90% מהנפח הנצרך בארצות הברית. מה שהופך את שוק האבוקדו למאוד רגיש לתנאי מזג אוויר קיצוניים ושריפות בקליפורניה.[1]

3) ישנם כ- 7 זנים עיקריים של אבוקדו בארה"ב, כאשר הזן המוביל הוא "הס". זן זה נחשב לנוח לגידול, בולט בחיי המדף הארוכים שלו וניתן לגדלו כמעט כל השנה.

4) כמות ייצור האבוקדו בארה"ב עומדת על כ- 146 טון אבוקדו בשנה [2] (USDA, איור 1), המהווה הכנסה שנתית של כ- 932 מיליון דולר.



איור 1 - כמויות הייצור (בירוק), הייבוא (באדום) והצריכה (בכחול) של אבוקדו לאדם

מבנה הנתונים

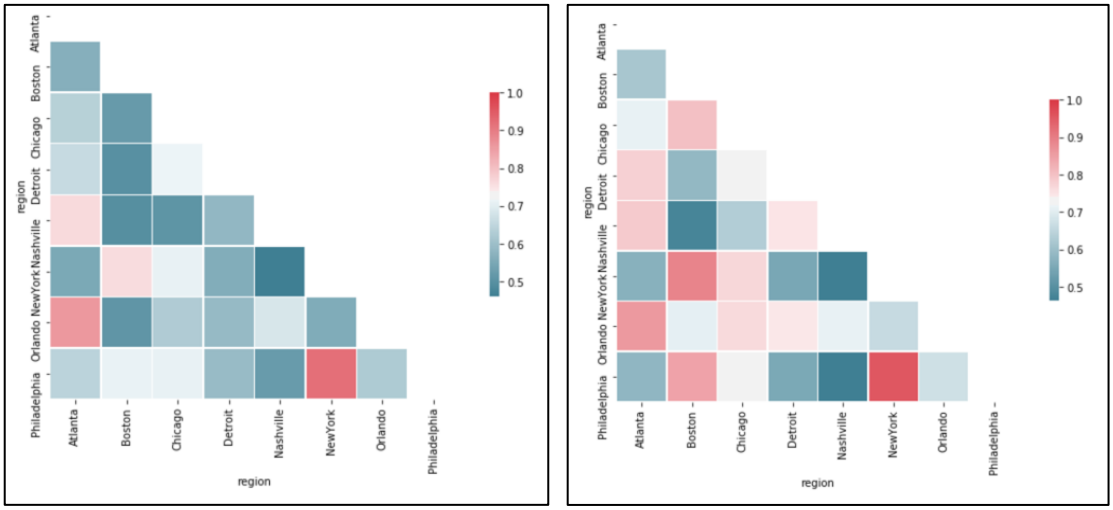
הנתונים הגולמיים שלנו (raw data) עוסקים במכירות ומחירי האבוקדו בארה"ב בין השנים 2015-2018.
מבנה הנתונים נלקח מאתר kaggle[3] וכולל את השדות הבאים:

- תאריך המכירה
- מחיר ממוצע (ליחידה)
- כמות האבוקדו שנמכרה, בחלוקה ל-3 גדלים (4046, 4225, 4770)
- סה"כ נפח אבוקדו נמכר
- כמות שקים שנמכרו (יחידות) בחלוקה ל-3 גדלים
- סה"כ שקים שנמכרו
- סוג הגידול (רגיל, אורגני)
- עיר

הכנת המידע ועיבוד מקדים (Data Preparation & Pre-Processing)

בחירת אזור המכירה הגאוגרפי

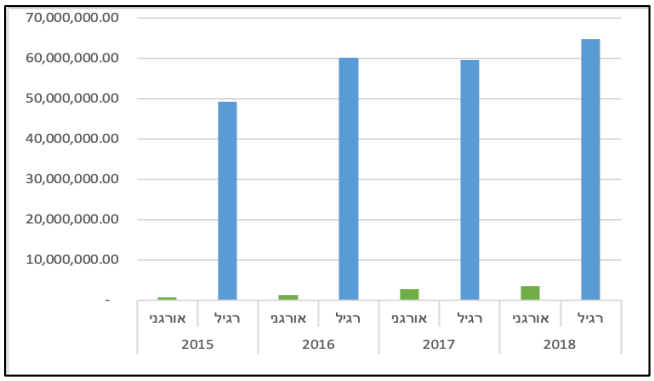
בסיס הנתונים הכיל כ-50 אזורי מכירה שונים. בחרנו להתמקד ב- 8 ערים בחוף המזרחי (אטלנטה, בוסטון, שיקגו, דטרויט, נאשוויל, ניו-יורק, אורלנדו ופילדלפיה), על מנת להראות בצורה נוחה את הפרמטרים וכיון שרצינו שניתוח הנתונים של המכירות יבוצע על אותו אזור.
מתוך בדיקות המתאמים (איור 2), עולה כי בחלק ניכר מהערים הנבחרות יש מתאם די חזק עבור המשתנה המסביר - המחיר הממוצע, וקיים מתאם (אם כי בצורה קצת פחות חזקה) עבור המשתנה המוסבר - נפח המכירה הכולל.
מכאן הסקנו כי יש כנראה משקל להתנהגות הצרכנית בכל עיר, ואזור גיאוגרפי דומה (החוף המזרחי) אינו מסביר את התופעה בצורה מספקת. ועל כן, כללנו את "עיר" כמשתנה מסביר במודל.



איור 2 - מטריצות קורלציה, ערים כפונקציה של מחיר ממוצע (מימין) וערים לפי נפח מכירה כולל (משמאל).

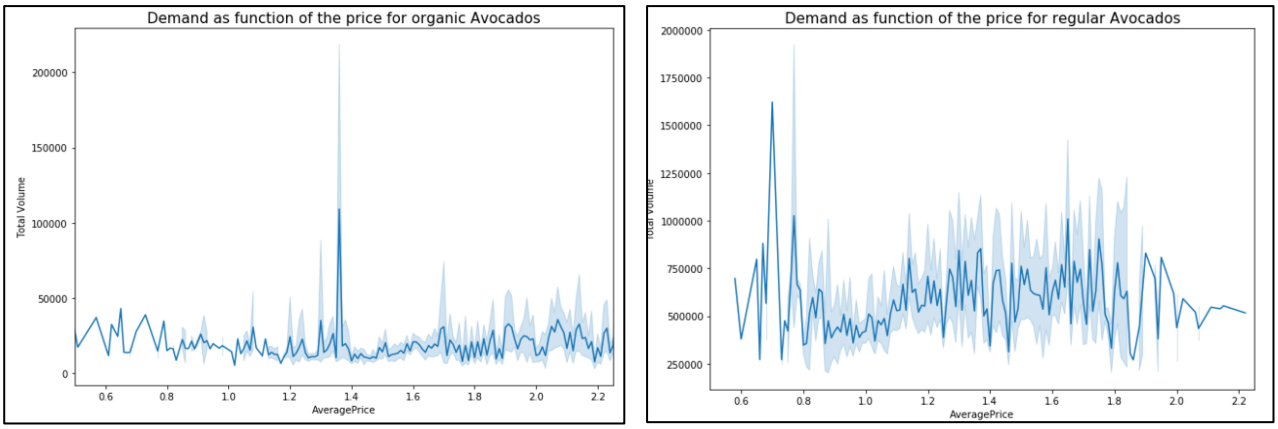
בחירת סוג האבוקדו והכמות

כפי שניתן לראות (איור 3), נפח המכירה של אבוקדו בגידול "אורגני" היה זניח יחסית לסוג הגידול "הרגיל" (בממוצע, פחות מ 3%).



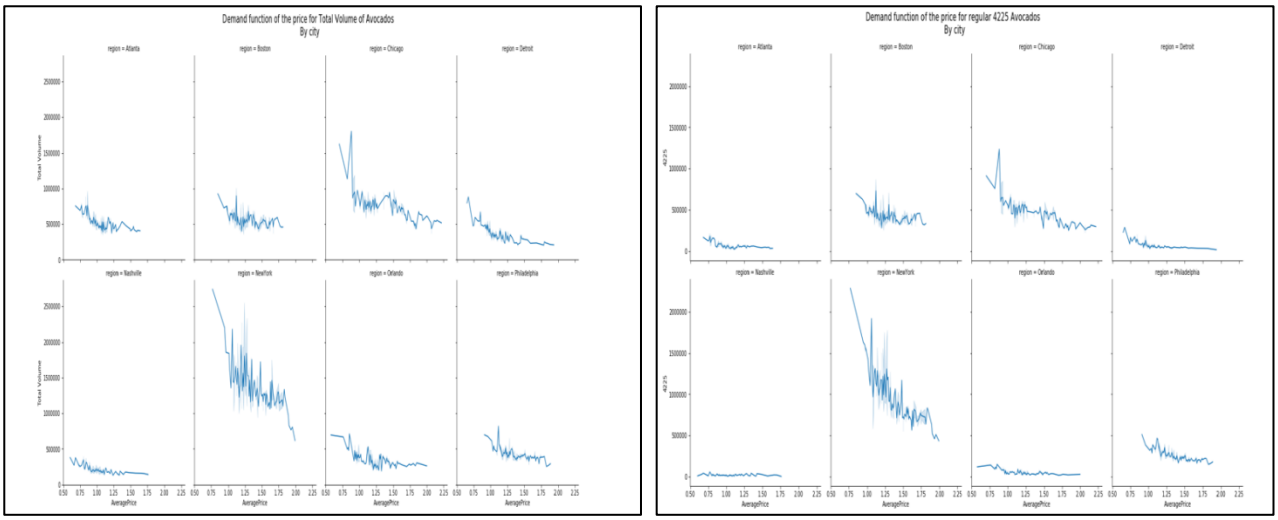
איור 3 - נפח מכירה כולל לפי סוג האבוקדו ולפי שנים.

בנוסף, התנודתיות שנצפתה במחיר של האבוקדו מהסוג "הרגיל" לעומת הסוג "האורגני", הייתה נראית הרבה יותר מעניינת לחקירה (איור 4), ולכן בחרנו להתמקד בעבודה זו בסוג האבוקדו "הרגיל".



איור 4 - גרף הביקוש כפונקציה של המחיר הממוצע, עבור הסוג "הרגיל" (מימין) ועבור הסוג "האורגני" (משמאל).

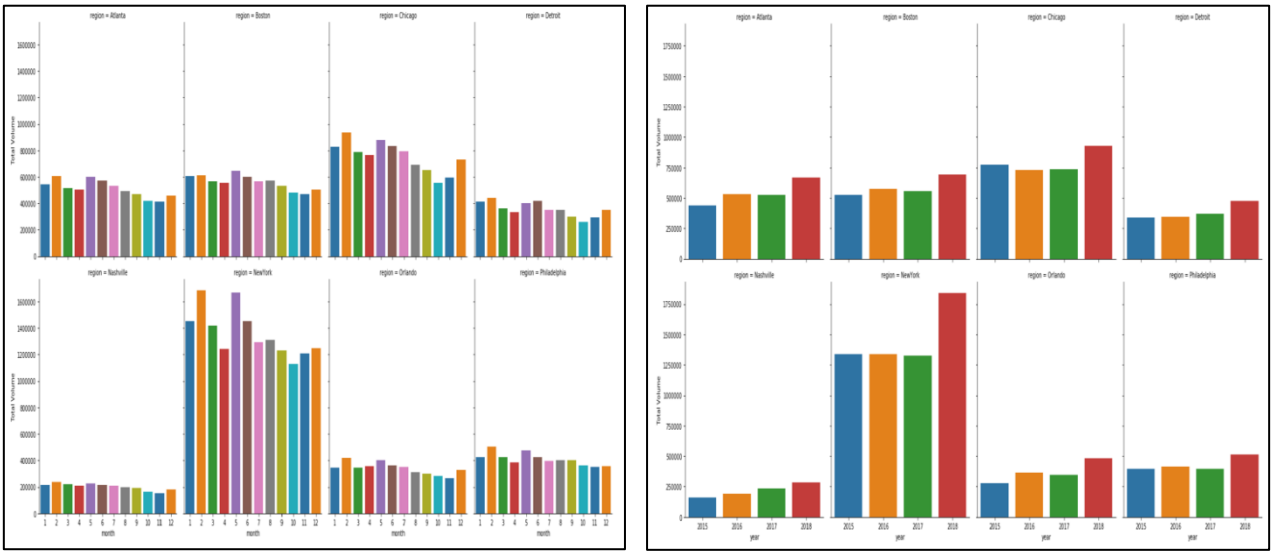
בשלב הבא, בדקנו האם להתמקד בכמות המכירה של אבוקדו מגודל 4225 או בנפח מכירה כולל (איור 5). כיוון שחלק מהערים אותן בחרנו, לא מכרו את הגדלים האחרים (4046, 4770) וכיוון שהתנודתיות במחיר עבור נפח המכירה הכולל נראתה הרבה יותר מעניינת לחקירה, החלטנו שהמודל יכלול את נפח המכירה הכולל (Total Volume) בתור המשתנה המוסבר.



איור 5 - גרף הביקוש בערים כפונקציה של המחיר הממוצע עבור סוג 4225 (מימין) ועבור נפח המכירה הכולל (משמאל).

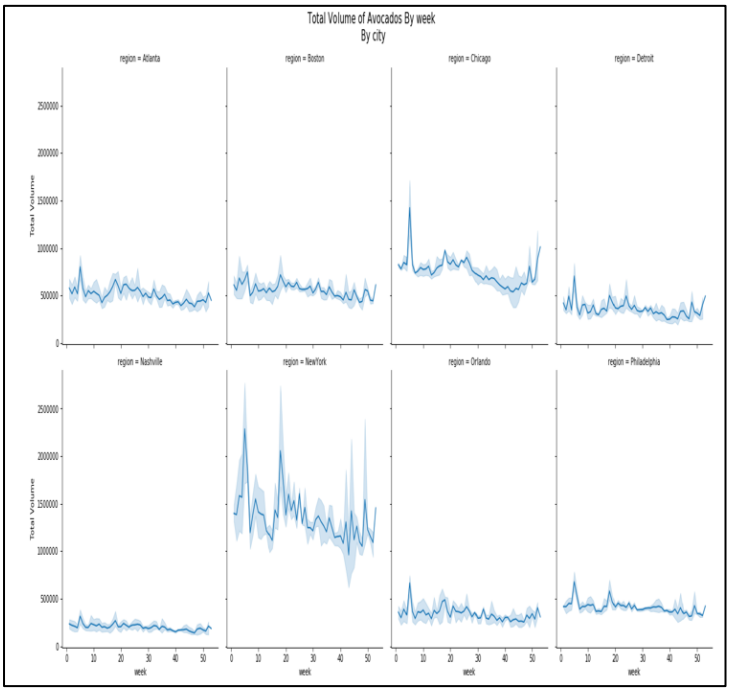
מגמות מכירה לאורך זמן

מחקירת נתונים נוספת, גילינו כי בכל הערים שבדקנו ישנה מגמת עלייה בצריכה הכוללת של אבוקדו במהלך השנים שנסקרו (איור 6).
יחד עם זאת, נתוני הצריכה לגבי החודש בשנה הראו עונתיות מסוימת כאשר החודשים פברואר ומאי הם החודשים המובילים בצריכה (איור 6). הסבר אפשרי לכך שמצאנו יכול להיות החג המקסיקני "ה-5 למאי" (Cinco de Mayo), בו קיימת צריכה מוגברת של ממרח הגוואקמולי העשוי מאבוקדו[4].



איור 6 - גרף הצריכה הכוללת בערים כפונקציה של שנת המכירה (מימין) וחודש המכירה בשנה

לבסוף, בדקנו האם קיימת עונתיות ברמה השבועית של נתוני המכירות לפי השבוע בשנה (איור 7).
הגרף הראה נתונים דומים כאשר בשבועות של פברואר (CW5-6) ומאי (CW19), ניתן לראות את הקפיצות הרלוונטיות בצריכה, אולם החלטנו בסופו של דבר לא להעמיק יתר על המידה בנושא העונתיות.



איור 7 - גרף הצריכה הכוללת בערים כפונקציה של שבוע המכירה בשנה.

מודל פונקציית הביקוש (Demand function model)

רגרסיה לינארית

המודל שבחרנו לבדוק על מנת לחזות את הביקוש היה רגרסיה לינארית.
כפי שהוסבר לעיל, המשתנים המסבירים במודל היו: **המחיר הממוצע (AvaragePrice)**, **הערים שבחרנו**, **חודש המכירה והשנה**.
כאשר חודש ושנה הומרו ממשתנים נומריים למשתנים בינאריים (one hot encoding).

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.872e+05	2.09e+04	42.489	0.000	8.46e+05	9.28e+05
AveragePrice	-4.586e+05	1.81e+04	-25.293	0.000	-4.94e+05	-4.23e+05
region_Boston	1.568e+05	1.33e+04	11.747	0.000	1.31e+05	1.83e+05
region_Chicago	3.849e+05	1.38e+04	27.950	0.000	3.58e+05	4.12e+05
region_Detroit	-1.28e+05	1.27e+04	-10.094	0.000	-1.53e+05	-1.03e+05
region_Nashville	-3.358e+05	1.27e+04	-26.471	0.000	-3.61e+05	-3.11e+05
region_NewYork	1.011e+06	1.4e+04	72.158	0.000	9.84e+05	1.04e+06
region_Orlando	-1.041e+05	1.29e+04	-8.061	0.000	-1.29e+05	-7.88e+04
region_Philadelphia	4.854e+04	1.4e+04	3.475	0.001	2.11e+04	7.59e+04
month_2	4.736e+04	1.42e+04	3.343	0.001	1.96e+04	7.51e+04
month_3	1.056e+04	1.4e+04	0.755	0.450	-1.69e+04	3.8e+04
month_4	2.941e+04	1.54e+04	1.907	0.057	-844.916	5.97e+04
month_5	1.218e+05	1.5e+04	8.120	0.000	9.24e+04	1.51e+05
month_6	9.292e+04	1.57e+04	5.901	0.000	6.2e+04	1.24e+05
month_7	6.616e+04	1.53e+04	4.338	0.000	3.62e+04	9.61e+04
month_8	5.582e+04	1.56e+04	3.573	0.000	2.52e+04	8.65e+04
month_9	4.89e+04	1.62e+04	3.021	0.003	1.72e+04	8.07e+04
month_10	3.248e+04	1.6e+04	2.027	0.043	1038.541	6.39e+04
month_11	-2.706e+04	1.55e+04	-1.742	0.082	-5.75e+04	3409.852
month_12	-4.922e+04	1.53e+04	-3.225	0.001	-7.92e+04	-1.93e+04
year_2016	5.428e+04	8133.332	6.673	0.000	3.83e+04	7.02e+04
year_2017	1.489e+05	9277.271	16.053	0.000	1.31e+05	1.67e+05
year_2018	2.516e+05	1.47e+04	17.114	0.000	2.23e+05	2.8e+05

Omnibus:	1088.820	Durbin-Watson:	1.348
Prob(Omnibus):	0.000	Jarque-Bera (JB):	56057.447
Skew:	3.315	Prob(JB):	0.00
Kurtosis:	33.841	Cond. No.	20.3

איור 8 - תוצאות הרגרסיה הלינארית וניתוח ה p-value עבור כל אחד מהמשתנים.

המשתנה המוסבר שלנו היה: **נפח המכירה הכולל** (Total Volume).
מדדי טיב הרגרסיה (איור 9), הראו כי אחוז השונות המוסברת במודל ע"י המשתנים שנבחרו היה גבוה (Adj. R-squared = 0.905).

OLS Regression Results

Dep. Variable:	Total Volume	R-squared:	0.907
Model:	OLS	Adj. R-squared:	0.905
Method:	Least Squares	F-statistic:	589.0
Date:	Tue, 18 Jun 2019	Prob (F-statistic):	0.00
Time:	22:06:05	Log-Likelihood:	-17676.
No. Observations:	1352	AIC:	3.540e+04
Df Residuals:	1329	BIC:	3.552e+04
Df Model:	22		
Covariance Type:	nonrobust		

איור 9 - מדדי טיב הרגרסיה הלינארית.

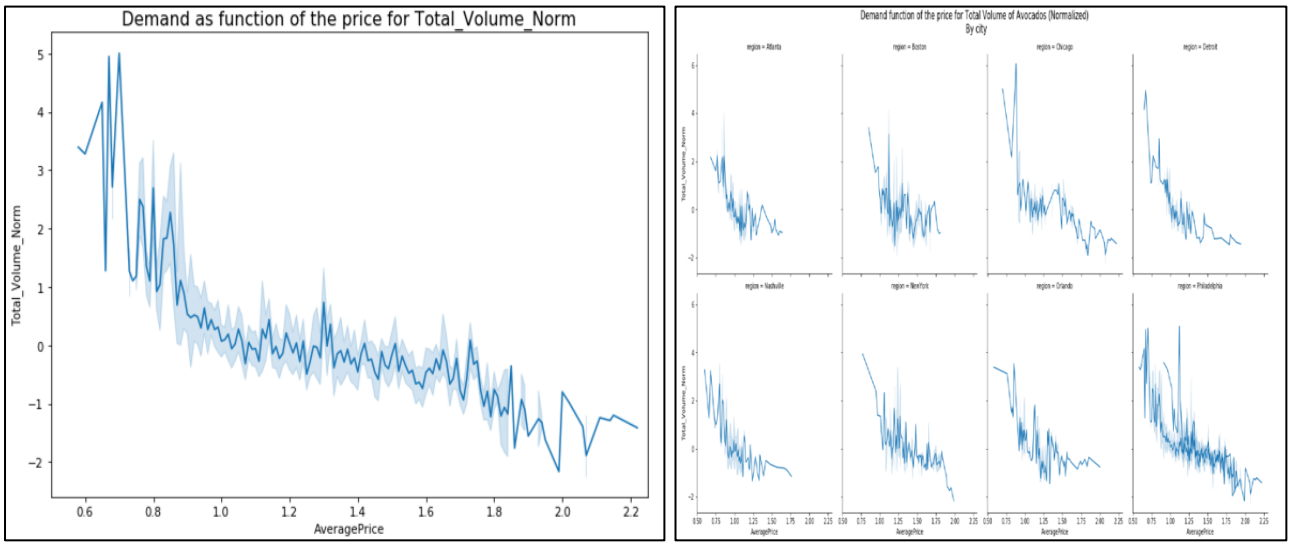
אומד המחיר האופטימלי

המשוואה הבאה, מייצגת את אומד המחיר האופטימלי, על סמך הרגרסיה הלינארית שהרצנו, עבור אותו סט של משתנים מסבירים שבחרנו.

$$\hat{p} = \frac{a + b_1 x_1 + b_2 x_2 + \cdots + b_{n-1} x_{n-1}}{2b_n}$$

גמישות הביקוש

לצורך חישוב גמישות הביקוש, נרמלנו (Z-Score normalization) את עמודת סה"כ נפח המכירה לפי עיר, כלומר, מהו סך כל הביקוש לאותה עיר (איור 10).
הנרמול היה כאשר המשתנה המסביר היה המחיר הממוצע, והמשתנה המוסבר היה נפח המכירה.



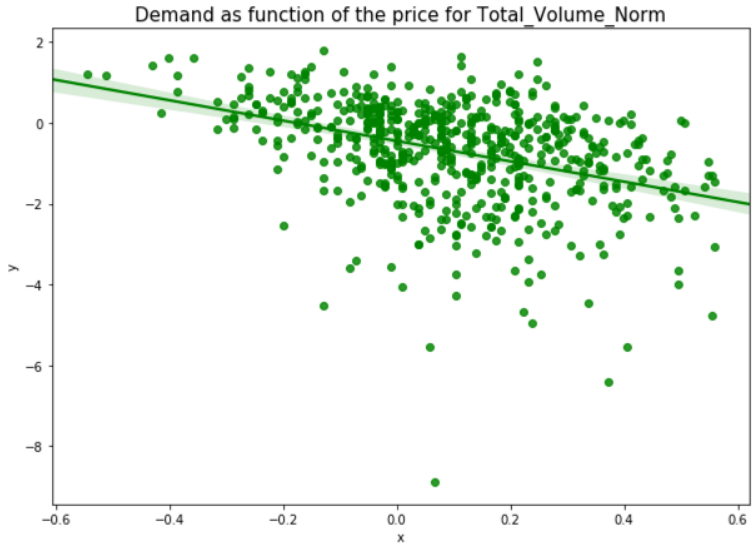
איור 10 - גרף הביקוש כפונקציה של המחיר עבור סה"כ נפח מכירה מנורמל לעיר (מימין), ועבור סה"כ נפח מכירה מנורמל כולל (משמאל).

מאחר ואנו מנסים לאמוד את אלפא המקיימת $x^{\alpha} = y$ נפעיל לוג על שני האגפים כדי שנוכל לאמוד את אלפא על ידי רגרסיה ליניארית -

$$\log(x^{\alpha}) = \log(y)$$

$$\alpha * x = \log(y)$$

לאחר הרצת הרגרסיה עם הערכים המנורמלים, (כ-30 תצפיות עבור כל פרמטר), ערך האלסטיות שהתקבל עבור המודל של כל הערים יחדיו **הוא 2.5** (ערך החותך ברגרסיה). דבר זה מעיד כי גמישות הביקוש לאבוקדו אינה קשיחה (שכן היא גדולה מ-1), וזאת בהתאם למוצר שאינו בסיסי (איור 11).



איור 11 - גרף הרגרסיה הלינארית, ביקוש כפונקציה של סה"כ נפח מכירה כולל מנורמל.

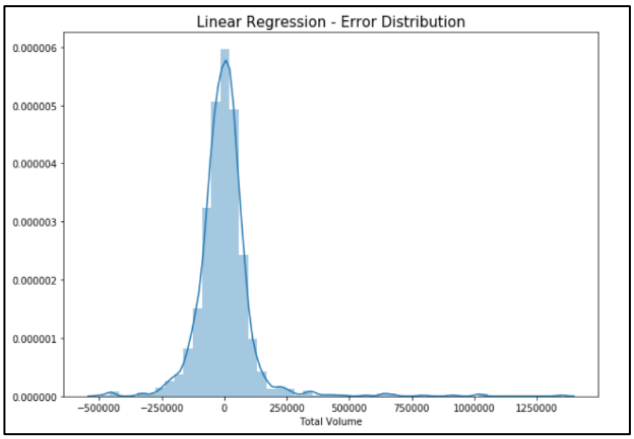
תוצאה זו אף עולה בקנה אחד עם מוצרים אחרים משוק הירקות שראינו בשיעור, כמו למשל אפונה טרייה (2.8) או עגבניות (4.6).
בנוסף, בחנו את גמישות הביקוש של כל אחת מהערים בנפרד, ונמצאו הבדלים משמעותיים בערך האלסטיות, כפי שניתן לראות בטבלה הבאה:

Atlanta	Boston	Chicago	Detroit	Nashville	NewYork	Orlando	Philadelphia
4.11363	2.56136	2.68093	6.76077	3.33371	4.34474	3.22169	5.63046

בהקשר זה, יהיה מעניין לחקור בעתיד את ההבדלים בין הערים ומה יכול להיות הגורם לשוני זה (האם זהו המצב הסוציואקונומי? ההכנסה הממוצעת למשפחה? וכו').

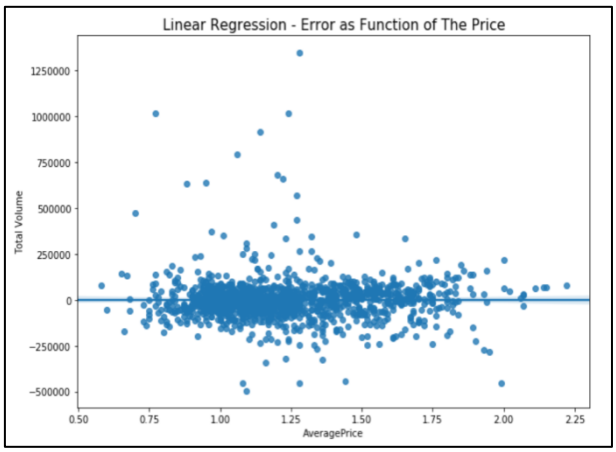
ניתוח השונות והפיזור במחירים הנצפים

על מנת לבחון את טיב המודל שבנינו, השתמשנו במדדי טיב התאמה של מודל ריבועים פחותים (OLS).
התפלגות השגיאות על פני המשתנים השונים הייתה טובה יחסית. עבור נפח כולל (איור 12), ניתן לראות כי השגיאות מתפלגות באופן נורמלי סביב הערך 0, עם זנב ימני ארוך יחסית (Right-skewed).

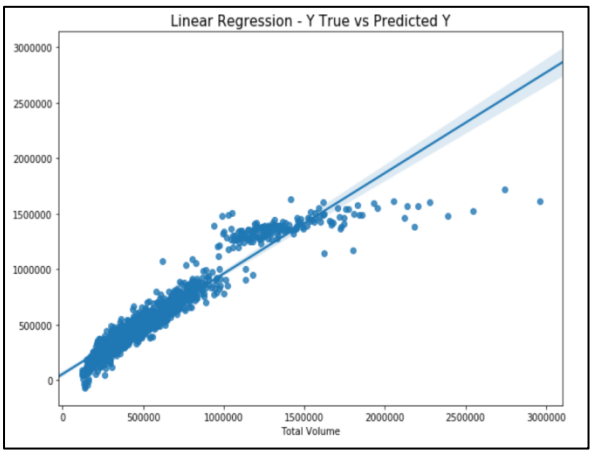


איור 12 - התפלגות השגיאות Error ~ Total Volume

באופן דומה, ניתן לראות בגרף התפלגות השגיאות עבור המחיר הממוצע (איור 13), ובגרף התפלגות השגיאות עבור נפח המכירה החזוי לעומת נפח המכירה האמיתי כי השגיאות אינן מפוזרות בצורה אחידה לחלוטין.
באחרון (איור 14), אף ניתן לראות בעיית התאמה בחלק העליון הימני של הגרף, וזאת כיוון שהמודל שלנו מנסה לחזות באותו אופן גם עבור מחירים נמוכים וגם עבור מחירים גבוהים.

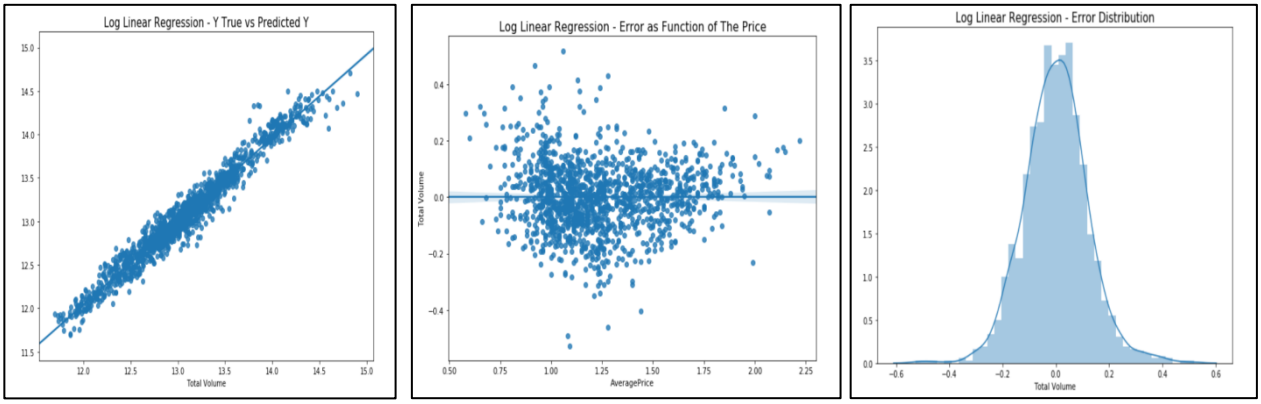


איור 13 - Error ~ Average Price



איור 14 - Total Volume (predicted) ~ Total Volume (true)

בגלל האמור לעיל, וכיוון שההתפלגות של הערך החזוי Y הינה בעלת זנב ימני ארוך (Positive Skewness), החלטנו לבחון את השגיאות במודל רגרסיה נוסף, **הרגרסיה הלוגיסטית** (Log Linear Regression).
כפי שניתן לראות מטה (איור 15), במודל זה הייתה התאמה טובה יותר והתפלגות השגיאות של הערך החזוי שהתקבלה הייתה הרבה יותר אחידה.



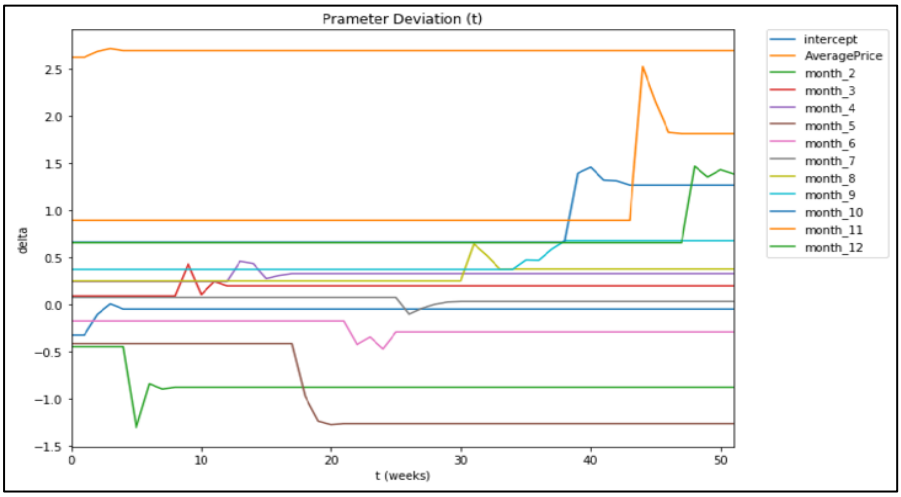
איור 15 - התפלגות השגיאות Error ~ Total Volume (מימין), Error ~ Average Price (באמצע), Total (predicted) ~ Total (true) (משמאל)

אולם בשלב זה החלטנו שלא לשנות את המודל איתו התחלנו.

הרצת המודל והערכת המחיר האופטימלי בזמן אמת

למידת הפרמטרים

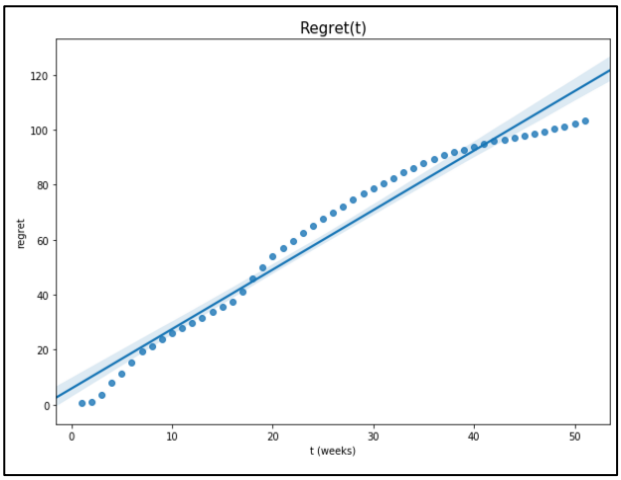
אמידת המחיר האופטימלי בזמן אמת היא פונקציה של הפרמטרים שנלמדו עד כה. במקרה שלנו, בו הביקוש תלוי בזמן (פרמטר החודש) ומספר התצפיות עבור כל חודש קטן (4 תצפיות לכל עיר בכל חודש) - נצפה שהפרמטרים לא יתכנסו לערכם 'האמיתי'.
בגרף הבא נדגים זאת על ידי סימולציה המדמה מצב אמיתי בו העסק פועל כשנה אחת בלי מידע מקדים על ערך הפרמטרים. בכל שבוע מתקבלת דגימה חדשה לכל עיר (8 ערים, 52 שבועות) - ניתן לראות שעבור כל פרמטר של חודש יש תקופה של 4 שבועות בו ערך הפרמטר מתעדכן ומתבצעת למידה כי רק שם קיימות תצפיות רלוונטיות לפרמטר זה (באופן דומה לעדכון בייסיאני).
הגרף מציג את הסטייה בין הערך הנאמד בזמן t, לבין הערך האמיתי כפי שנלמד בהינתן כל הנתונים.
נשים לב שאכן הפרמטרים לא מתכנסים כפי שצפינו.



איור 16 - למידת הפרמטרים (month) כתלות בזמן t (שבועות) עבור שנה אחת.

פונקציית החרטה (Regret) בתלות הזמן

פונקציית החרטה מציגה לנו את ההפסד הכספי כתלות בזמן הנובע מסטוכסטיות אמידת הפרמטרים.
כפי שניתן לראות (איור 17), קיבלנו גרף ליניארי האומר שהחרטה היא קבועה ליחידת זמן. הקו הישר בגרף הוא קו המגמה (רגרסיה) המספק אמדים לשיפוע ולחותך של הישר.
בנוסף, הגרף (איור 18), מראה את ההבדל בין המחיר ה'אופטימלי' לעומת המחיר הנאמד. נשים לב שבאופן כמעט קבוע המחיר הנאמד נמוך מהאופטימלי, נתון אשר פותח צוהר למחקר נוסף. נציין שמאחר ומדובר בסימולציה אחת בלבד, זהו לא ממצא סטטיסטי ולא וידאנו זאת באמצעות הרצת ניסויים נוספים.

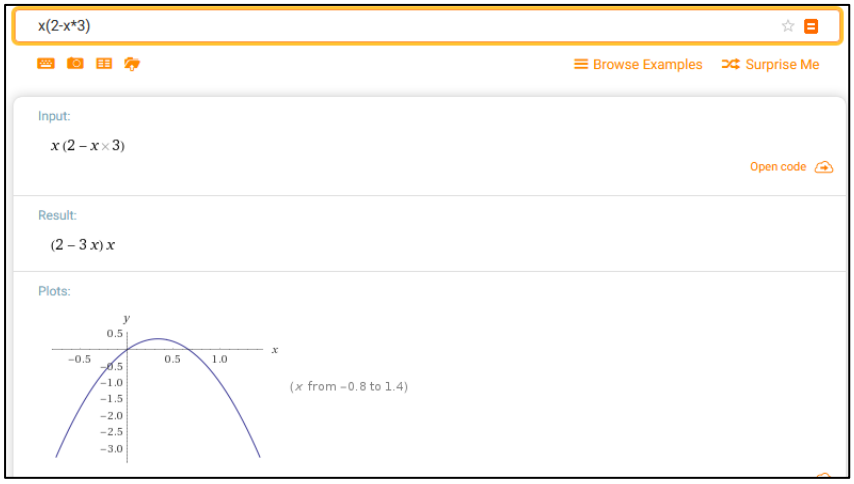


איור 17 - פונקציית החרטה כתלות בזמן t (שבועות) עבור שנה אחת.

איור 18 - מחיר כפונקציה של זמן t (שבועות) עבור שנה אחת.

דבר נוסף אותו נרצה לציין הוא שבהמשך להערה מהמצגת על מטריקות טובות יותר למדידת המודל, שמנו לב שבמקרה של מידול פונקציית הביקוש כפונקציה ליניארית, פונקציית הרווח המתקבלת היא פונקציה ריבועית (פרבולה). כלומר, סטייה קטנה סביב המינימום תוביל לסטייה קטנה יותר ברווח.
משמעות הדבר היא שחשוב יותר להגיע לסביבת המינימום ופחות לנקודה הקריטית.

הגרף הבא (איור 19) ממחיש דוגמה פרטית ומנוונת לעקרון זה:



איור 19 - פונקציית רווח שטוחה סביב המינימום בהנחת מודל ביקוש ליניארי

תובנות, מסקנות וכיווני מחקר נוספים

תובנות ומסקנות

- כפי שראינו מניתוח הנתונים, הביקוש לאבוקדו בערים שנבדקו גדל עם הזמן. כמו כן, מצאנו כי אבוקדו הוא מוצר "מותרות" בעל גמישות ביקוש הגדולה מ-1 ("ביקוש גמיש").

- התנהגות הצריכה של אבוקדו שונה מעיר לעיר בארה"ב, דבר המחזק את הטענה כי אבוקדו אינו מוצר בסיס אלא מוצר "מותרות" עם גמישות ביקוש המותאמת כנראה למצב הסוציואקונומי באותה עיר.

- אבוקדו נצרך בצורה שונה בהתאם לעונות, וזאת למרות שהוא פרי הנחשב כזמין בכל עונות השנה בחוף המזרחי של ארה"ב. השערתנו היא שסיבות לכך יכולות להיות חג מקסיקני שבו צורכים כמות גדולה של גוואקמולי המיוצר מאבוקדו.

- מחירי האבוקדו בארה"ב נעים בין קצת פחות מ-1 דולר ל-3 דולר בתקופות קשות בהן היו אירועי בצורת ושריפה של שטחים חקלאיים (כפי שתואר במבוא), כאשר ב-2018 המחיר הממוצע של אבוקדו היה כדולר וחצי.
ניכר כי זו סטייה משמעותית מתוצרי הסימולציה שהרצנו (איור 18), דבר אשר יכול לרמוז (בהנחה כמובן שסוחרי האבוקדו יודעים פחות או יותר את מחירי האבוקדו), שהמודל שלנו לא מכויל היטב - ע"י כך שנקטנו בגישה פשטנית מדי בחיזוי הביקוש (הנחת ליניאריות) או חוסר במשתנים מסבירים.

כיווני מחקר נוספים

כיווני מחקר נוספים עליהם חשבנו ויכולים להשלים את התמונה הכוללת:

- הוספת נתונים דמוגרפיים (מוצא, מס' נפשות, העדפה תזונתית וכו') וסוציואקונומיים (כגון הכנסה למשק בית) של הרכב האוכלוסייה בערים הנדגמות, וזאת על מנת להצליב ולראות אם זה שוק שניתן "לסחוט" בו עוד את המחיר. ניתן לחשוב על מודלים שונים לשכונות שונות בתוך אותה העיר.

- השוואה מול שווקים נוספים בארה"ב (West Coast), וזאת על מנת לקבל תמונה יותר כוללת של השוק האמריקאי.

- היצע אבוקדו - הצלבת הנתונים עם מידע (data set) נוסף של היצע האבוקדו באותו שוק נבחר, וזאת בכדי לבחון את ההשפעה על המחירים.

- הוספת פרמטרים נוספים למודל והרצה של מודלים מורכבים יותר - למידה ממושכת, ניתוחי עונתיות נוספים סביב תאריכים מיוחדים שהוזכרו, שימוש במודלים אחרים שהוזכרו כגון מודלים אוטורגרסיביים, SVM וכו'.

- בחינת השפעות של אירועים חריגים (כגון: בצורת קשה בקליפורניה, ייבוא אבוקדו ממקסיקו וכו') על מחירי האבוקדו ונפח המכירות הכולל.

מקורות וביבליוגרפיה

[1] Mark A. The United States Avocado Market, http://www.avocadosource.com/temp/OLD%20WAC%20II/WAC2_p643.htm

[2] USDA (2017) Avocado imports play a significant role in meeting growing U.S. demand, https://www.ers.usda.gov/data-products/chart-gallery/gallery/chart-detail/?chartId=83396

[3] Dataset: Avocado Prices - Historical data on avocado prices and sales volume in multiple US markets, https://www.kaggle.com/neuromusic/avocado-prices

[4] California avocado, https://www.californiaavocado.com/retail/avocado-plus