Machine learning

Final project

המשימה:

במערך הנתונים, תקבלו נתונים של חנויות שונות של חברת סופרמרקט לפי מזהה החנות שלהם, שהומרו למספרים שלמים חיוביים.

מזהה חנות: מזהה (אינדקס) של החנות הספציפית.

Store\_Area: שטח פיזי של החנות בכיכר החצר.

Items\_Available: מספר פריטים שונים הזמינים בחנות המתאימה.

DailyCustomerCount: מספר הלקוחות שביקרו בחנויות בממוצע על פני חודש.

Store\_Sales: מכירות ב-(US $) שהחנויות ביצעו.

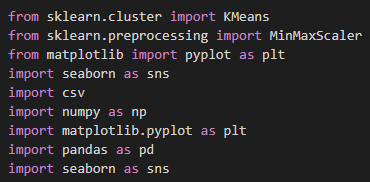
עליך בנות אלגוריתם למידה שמנבה את מחירות של חנות ספציפית

**תיאור הפרוייקט:**

בפרוייקט זה אנו רוצים לחזות את המחירות של חנות מסויימת לפי פרמטרים כמו : הגודל של החנות, המוצרים שזמינים בחנות ומספר לקוחות יומי.

בשלב הראשון נבדוק את הdataset שלנו שאין בו שורות שבורות( עמודות ריקות וכו..) ונבדוק שאין לנו משתנים שאינם מספרים, אם יהיו משתנים כאלה נהפוך אותם למספרים.

לאחר מכן נחפש קשרים בין המשתנים השונים שבעזרתם נצליח לנבה את המחירות של החנות.

הספריות שעשינו בהם שימוש:

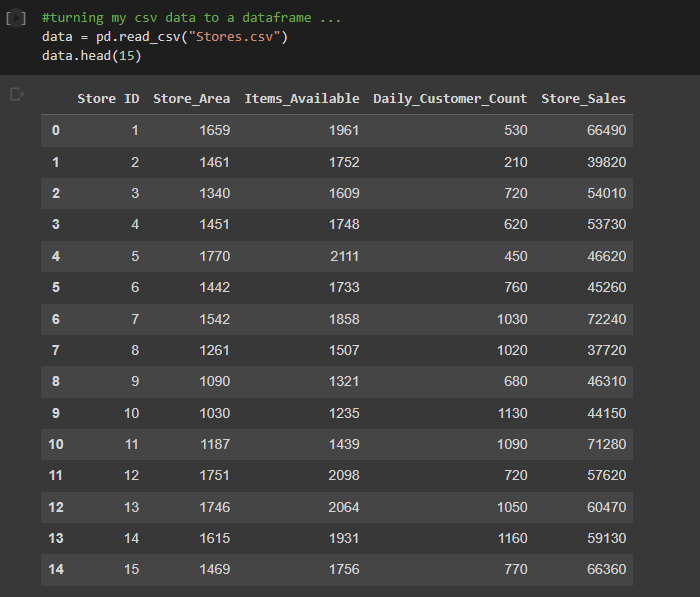
בחרנו להשתמש בספריה sklearn כפי

שהמרצה המליץ לצורך חיזוי , בנוסף, השתמשנו בספריות כמו seaborn לצרוך הצגת גרפים בצורה קלה יותר.

Pandas לפתיח קבצי csv.

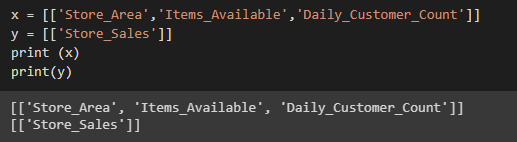
Numpy לחישובים מתמטים.

דבר ראשון לפני שמתחילים... נפתח את הdataset שלנו ונבין עם מה אנחנו עובדים.

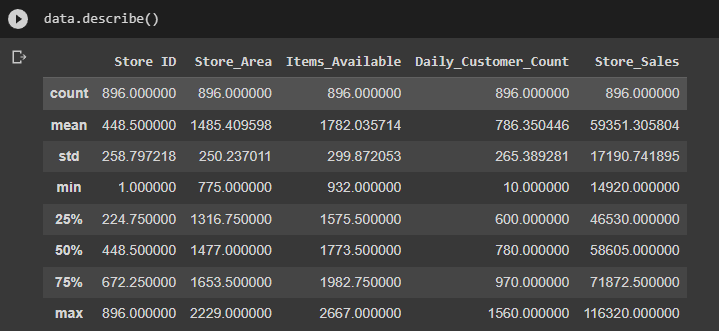


מצויין! כפי שניתן לראות, אין לנו עמודות שבורות או עמודות של אותיות שצריך להמיר אותן, ולכן לא צריך לבצע שום מניפולציה על המידע שלנו.

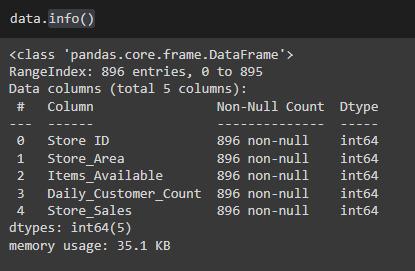
בשלב הבא נפריד את ה dataset למידע וחיזוי .

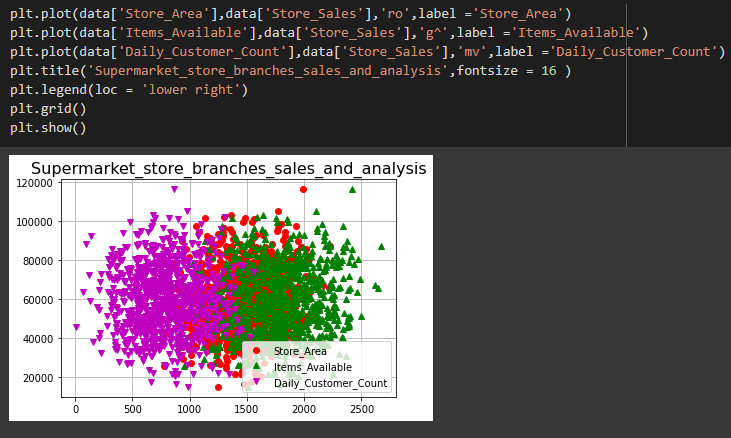


עכשיו נחקור קצת את המידע שלנו , נשתמש פונקצית describe() , בעזרת ניתן לראות את כמות השורות , את הממוצע, סייטת התקן , ערך מקסימאלי\מינימאלי , 50%,25%,75%



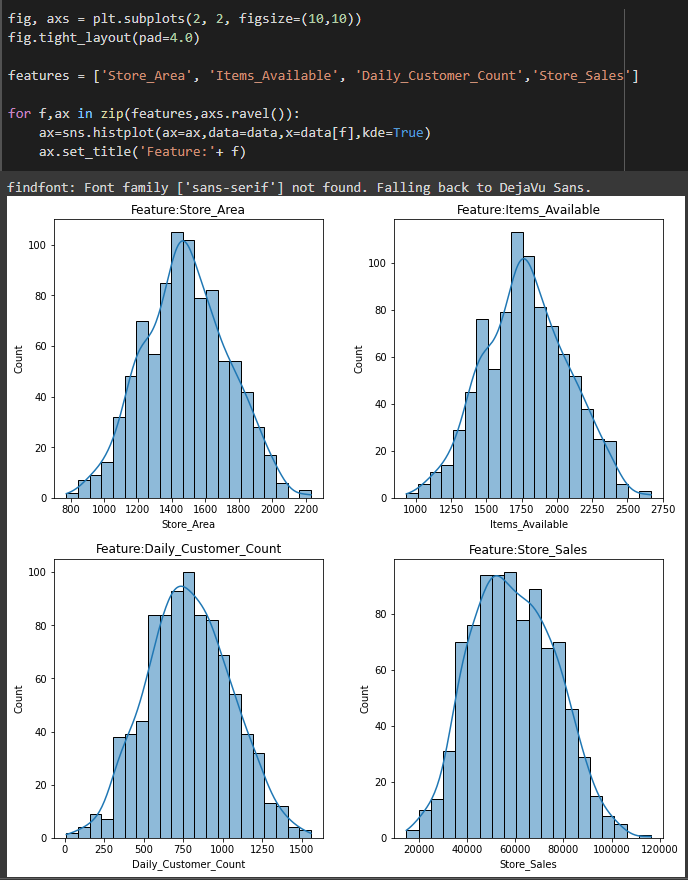
נעזר גם בפונקציה info() , כדי לקבל תמונה כללית על המידע שלנו , סוגי אינדקסים , עמודות ושימוש בזיכרון.



עכשיו נחפש קשרים לינארים בין המשתנים למכירות..

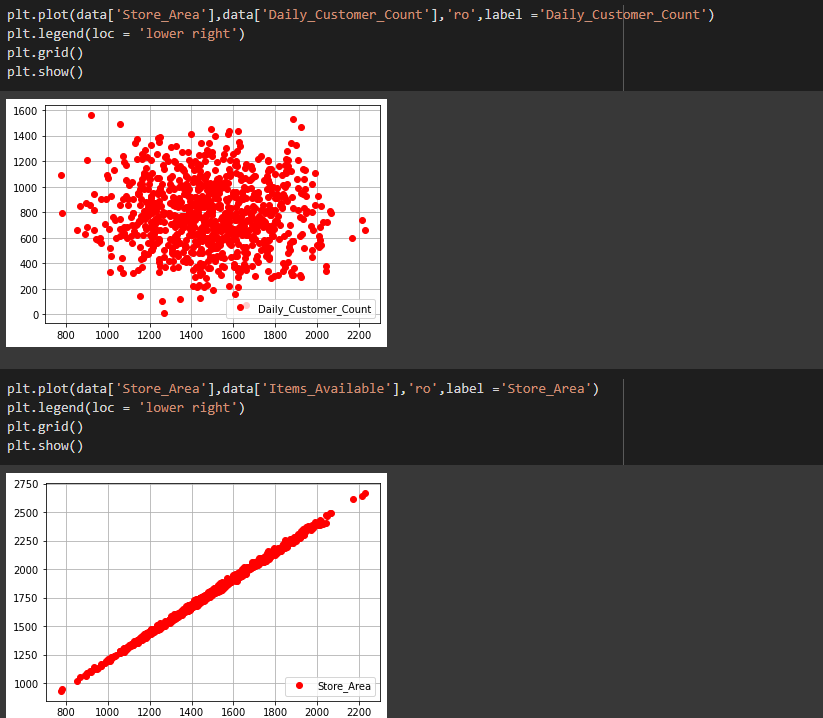
כפי שניתן לראות אין משתנה שקשור בצורה לינארית לפרמטר החיזוי שלנו...

ולכן נחקור קצת יותר את הפרמטרים שלנו ונבין יותר טוב את הפיצרים.



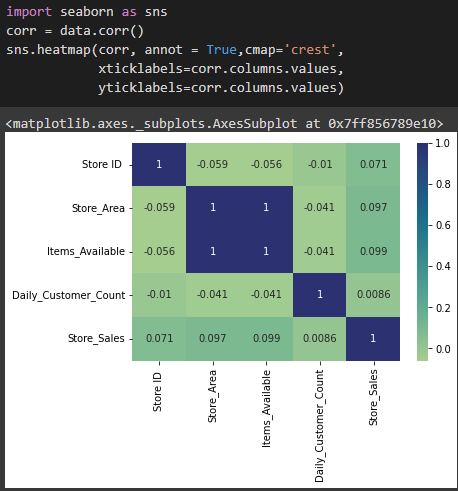
בתמונה הזאת ניתן לראות שכל הפיצרים שלנו מתנהגים בצורה סטטיסטית ולכולם יש התפלגות נורמאלית.

נחפש אולי עוד קשרים לינארים בין המשתנים..



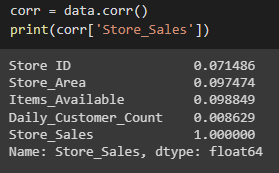
כמו שניתן לקרות, יש לנו קשר לינארי בין המשתנים: גודל החנות והמוצרים שזמינים, דבר שהוא מאוד הגיוני.

כדי לבדוק שאנחנו לא מפספסים שום קשר לינארי בין המשתנים נעזר בטבלת חום.

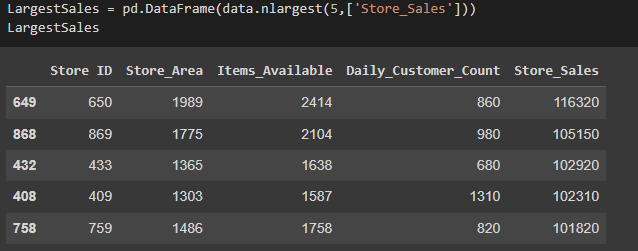


כלי מצויין לבדיקה הזאת היא corr(), אשר מציגה את כל הקשרים של המשתנים השונים בטבלת חום כללית.

ניתן גם להציג בטבלה רגילה.

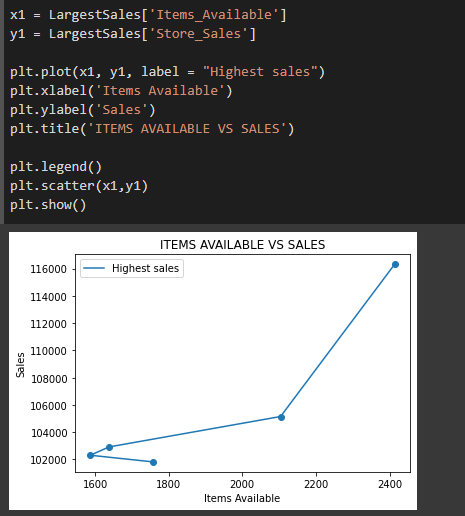


השלב הבא , אנו נרצה להוציא מהמשוואה את כל ה outliners , בעצם את כל המשתנים הגדולים מאוד והקטנים מאוד כדי שלא יפריעו לנו בחיזוי ולכן נשתמש בפונרציות הבאות: nlaragest() smallest()



A screenshot of a computer screen

Description automatically generated with medium confidence

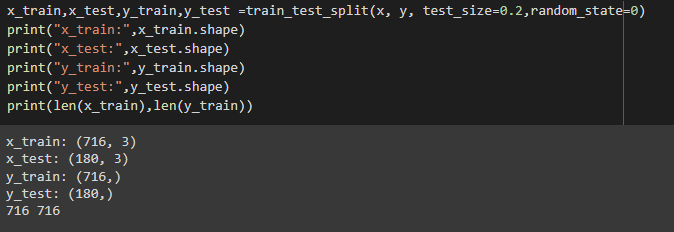
ניתן לראות מהקורלציה של מספר הפריטים בחנות משפיע על המכירות של החנות ( דבר הגיוני) יותר פריטים יותר מכירות! לכן נציג את ה 5 חנויות עם המכירות הגבוהות ביותר מול ה-5 חניות עם הכמות פריטים הגבוהה ביותר.

Chart, line chart

Description automatically generated

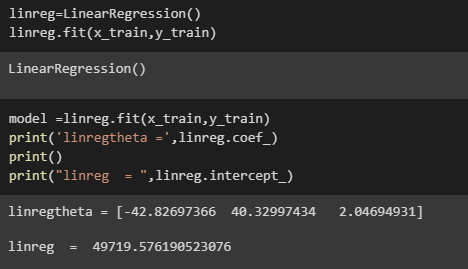
ועכשיו אחרי שבדקנו את המידע מכל הכיוונים נתחיל באימון המודל.

ראשית נחלק את המידע שלנו ל train & test , כמו ברוב המודלים , נחלק את המידע 20% test ו 80% train.



עכשיו נאמן את המודל בעזרת פונקציה fit(), הפונקציה מאמנת את האלגוריתם על ה train data.

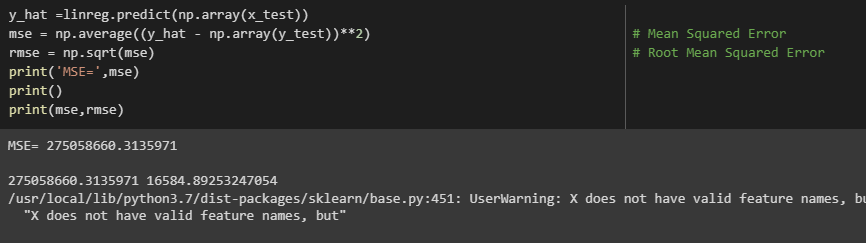
אז שיטת sklearn fit משתמשת בנתוני האימון כקלט כדי לאמן את מודל למידת המכונה.



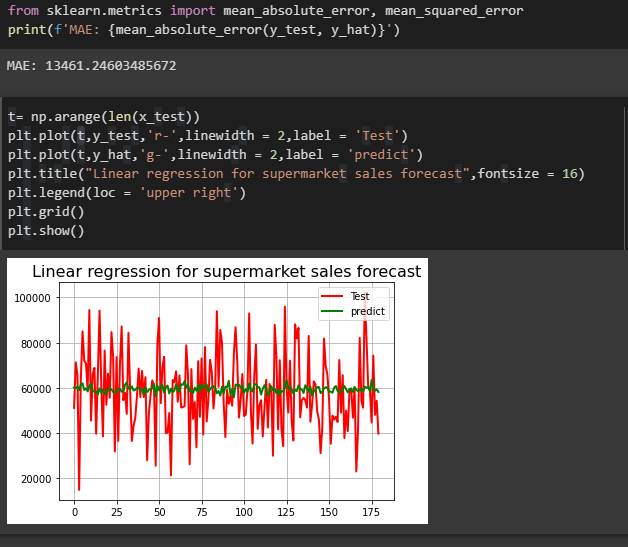
ניתן לראות שאילו המקדמים שקבלנו מפונקציה coef(), ואת הריגרסיה הליניארית.

הריגרסיה הלינארית במודל ריגרסיה מייצג את הערך הממוצע של משתני התגובה.

ועכשיו נמצא את ה mse, rmse.



Text

Description automatically generated

ניתן לראות שהגענו לאחוזי חזוי של 98% אחוזים!!!