Machine learning

Final project

המשימה:

חברת רכב שואפת להיכנס לשוק האמריקאי על ידי הקמת יחידת הייצור שלה שם וייצור מכוניות מקומית כדי לתת תחרות לעמיתיהם בארה"ב ובאירופה

הם התקשרו עם חברת ייעוץ לרכב כדי להבין את הגורמים שבהם תלוי תמחור המכוניות. באופן ספציפי, הם רוצים להבין את הגורמים המשפיעים על תמחור המכוניות בשוק האמריקאי, מכיוון שאלו עשויים להיות שונים מאוד מהשוק שהם מכירים. החברה רוצה לדעת:

אילו משתנים משמעותיים בחיזוי מחיר המכונית

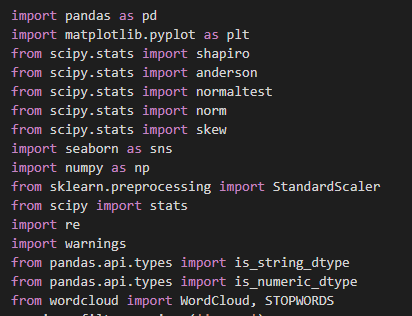
עד כמה משתנים אלה מתארים את מחיר המכונית

בהתבסס על סקרי שוק שונים, חברת הייעוץ אספה מערך נתונים גדול של סוגים שונים של מכוניות ברחבי השוק האמריקאי.

מטרה עסקית

אנו נדרשים לחזות את מחיר המכוניות עם המשתנים הבלתי תלויים הזמינים. זה ישמש את ההנהלה כדי להבין איך בדיוק המחירים משתנים עם המשתנים הבלתי תלויים. הם יכולים בהתאם לתמרן עם עיצוב המכוניות, האסטרטגיה העסקית וכו' כדי לעמוד ברמות מחיר מסוימות. יתר על כן, המודל יהווה דרך טובה להנהלה להבין את דינמיקת התמחור של שוק חדש

ספריות שנעבור איתם בפרוייקט:



בצגת ה dataset:

Graphical user interface, application

Description automatically generated

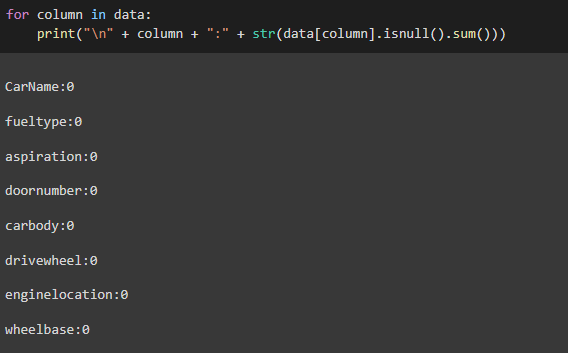
Text

Description automatically generated

נבדוק אילו משתנים יש לנו ב dataset :

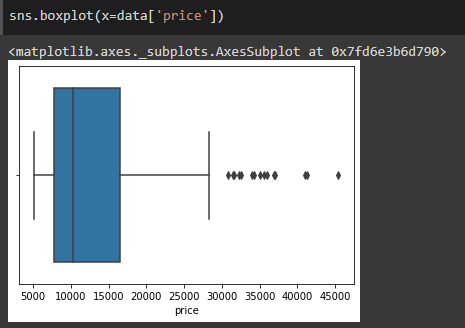
מהתבוננת במידע ניתן לראות שאין צורך במשתנה car\_ID ובמשתנה symbolling לחיזוי , ולכן נוציא אותם מהטבלה.

נבדוק שאין לנו עמודות ריקות במידע ואם כן נוריד גם אותם .

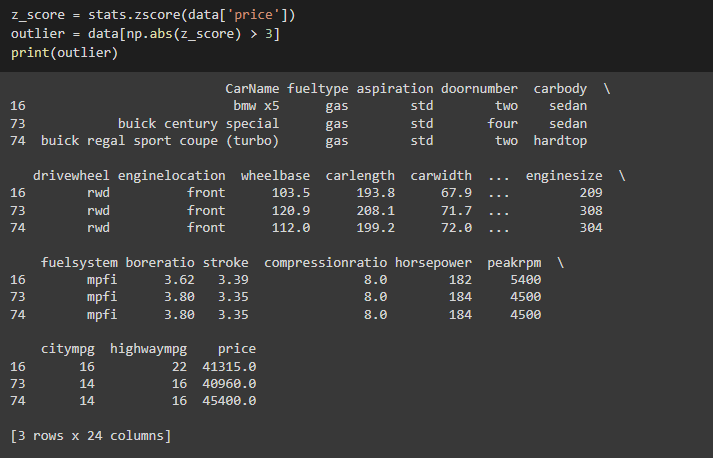


ניתן לראות שאין לנו שורות ריקות ומידע שלנו שלם וניתן לעבוד איתו.

לפני שנתחיל לחפש קשרים בין המשתנים שלנו נסתכל על המשתנה שאנחנו מחפשים : Price, נבדוק שאין לנו outliners. נרצה שלא יהיו חריגות ושההתפלגות תהיי נורמאלית בערך.

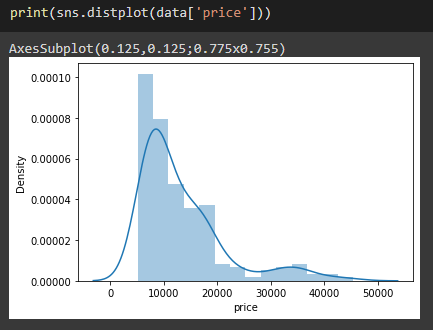


זה האמת ניראה שיש לנו תוצאות חריגות ולכן נסתכל על התוצאות החריגות ונבדוק האם זה מידע טוב או שניתן להוציא אותו.

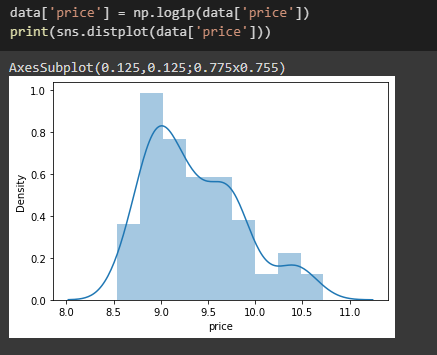


ניתן לראות שיש לנו 3 שורות שנחשבות outliers , אילו שלושה רכבים יקרים מאוד , יש לנו פה Buick 2 ואחד bmw, אילו באמת רכבים מיוחד והמחיר שלהם האמת גבוה, אנחנו נשאר עם המידע הזה , הוא לא באמת ישפיע לנו על המודל.

ננתח את המידע שלנו שלנו על המחיר ונבדוק האם הוא מתפלג נורמאלי.

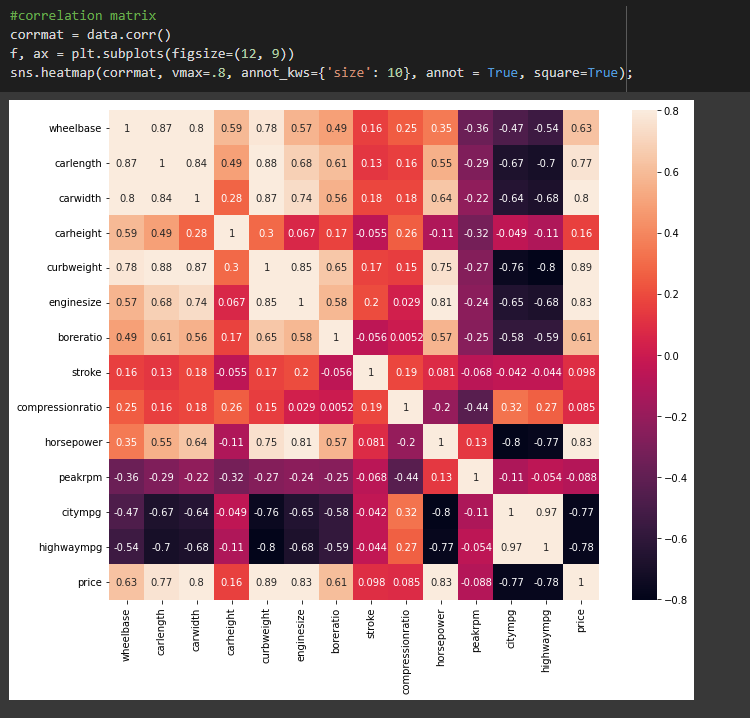


ניתן לראות שהתפלגות טיפה מוזזת , ניתן לסדר את זה בעזרת log-transformation.



עכשיו התפלגות המחירים ניראת הרבה יותר טוב.

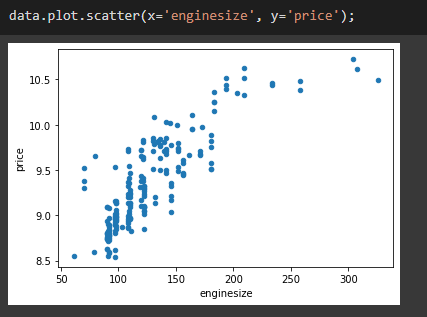
בשלב הבא אנו נרצה לחקור את הקשרים הליניארים בין המשתנים בdatase, נעזר בטבלת חום כדי להבין את קשרים אילו.



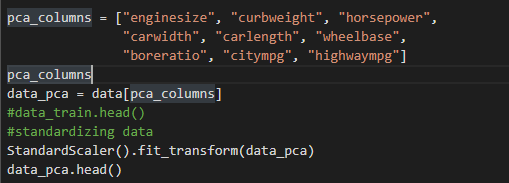
מהמפה ניתן לראות שלמחיר יש קשר לינארי חזק עם המשתנים הבאים:

enginesize, curbweight, horsepower, carwidth, carlength, wheelbase and boreratio.

ניתן לראות שיש גם קורלציה שלילית בין המחיר ל citympg and highwaympg , כאשר יש להם בינהם קשר ליניארי חזק.

נתחיל לחזור את המידע לפי המפה שמצאנו , אז תחילה נציג את המחיר למול גודל המנוע.

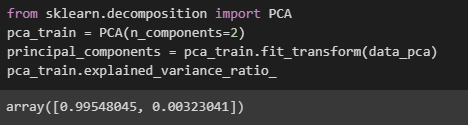
ממש ניתן לראות קשר לניארי חזק מבין המשתנים האילו, ניתן לראות שניתן להשתמש ברגרסיה לינארית.

מהניתוח הקודם ניתן להבין שיש משתנים רבים שיש להם קשר לניארי עם המחיר, ננסה להפחית את כמות המנבאים בעזרת טכניקת principle component analysis, כדי להשתמש בטכניקה זו ראשית אנו צריכים להנרמאל את המידע.

A screen shot of a computer

Description automatically generated with low confidence

עכשיו נמצא את 2 המספרים ליצירת הקו הישר לצורך יצירת קו ריגרסיה.

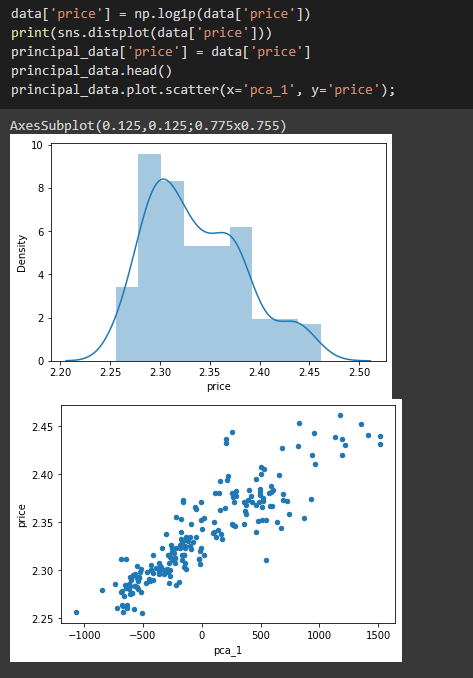


Graphical user interface, text

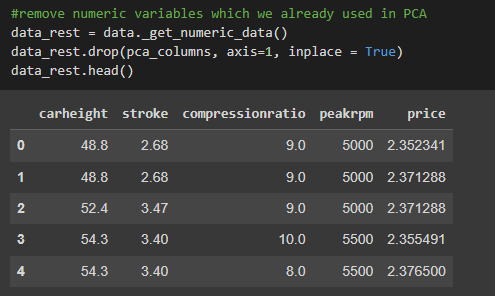
Description automatically generated

PCA מאפשר לנו להשתמש במשתנה אחד (מרכיב עיקרי) במקום 9 משתנים מספריים שהיו בקורלציה למחיר. בפעולה זו נאבד רק 0.3% מהמידע הכלול במשתנים זה.

נציג שוב את הגרפים:



לפי המפת הקורלציה ראינו אילו משתנים יש להם קשר לינארי עם מחיר, אבל מה עם המשתנים האחרים? האם יש להם קשרים לא לינארים עם המחיר ונדוק את זה עכשיו.



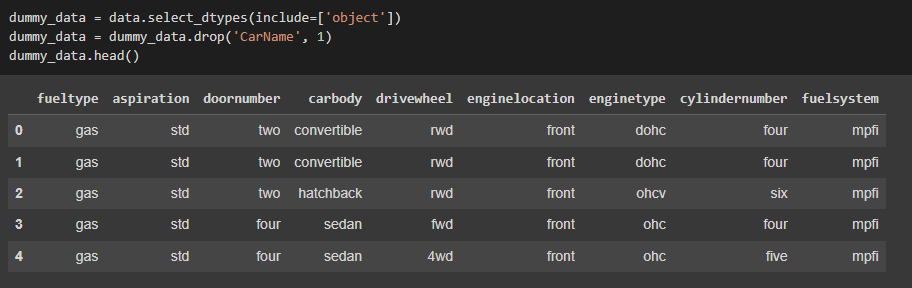
A picture containing diagram

Description automatically generated

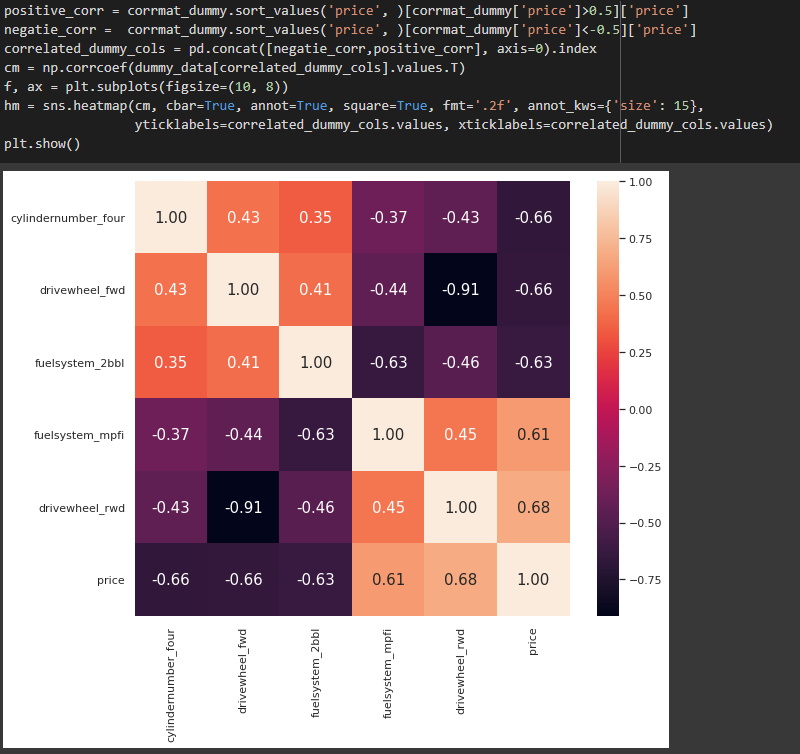
אנו לא רואים תבנית שיכולה לרמז לנו על קשרים בין המשתנים השונים למחיר הרכב.

לכן נמשיך לעבוד על המשתנים שכבר מצאנו , נבדוק אילו משתנים הכי מתאימים לחיזוי.

נבדוק אילו שורות יש להם משתנים שהם לא מספרים ונוציא ואתם מהטבלה שלנו.



כמו כן ניצור להם משתני דמה כדי להבין אילו משתנים אנו רוצים לקחת לחיזוי.

כעת נציג שוב את המפת הקורלציה רק שהפעם נציג רק הערכים אשר נמצאים בין 0.5 ל-0.5- .

זה ניראה שלמחיר יש קשרים חזקים עם המשתנים שראינו מקודם..

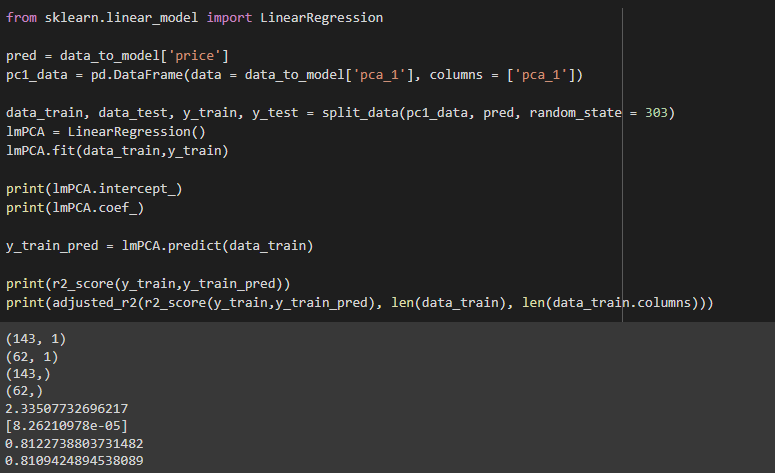
כעת ניתן להתחיל לאמן את המודל שלנו , בשלב הראשון אנו צריך לפצל את המידע שלנו ל train & test .

אנחנו החרנו לקחת 30% לtest ו70% train.

Text

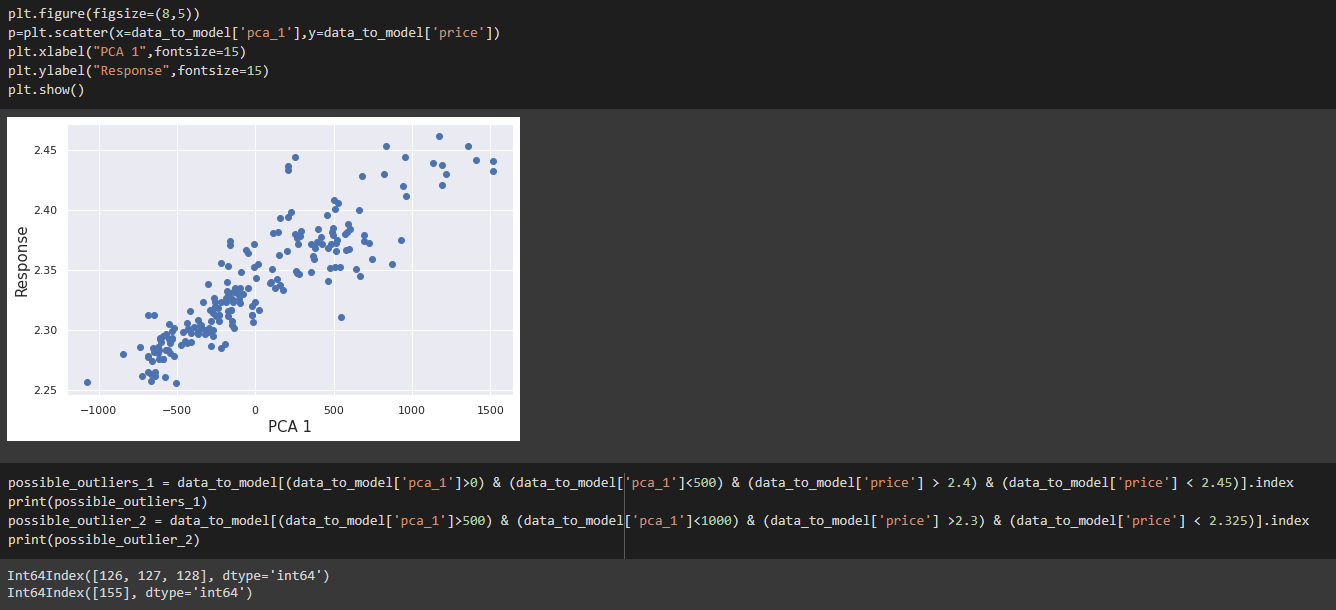
Description automatically generated

נתחיל לבנות את המודל , נתחיל לבשלב הראשון לנבות בעזרת PCA, שיצרנו בתחילת התוכנית וניראה אם אנחנו מקבלים תוצאות טובות.



ניתן לראות שמודל הזה מצליח לחזות ב 81% , שאילו אחוזים לא רעים בכלל.

נבדוק אם בעזרת ריגרסיה לינארית יהיה ניתן לקבל תוצאות טובות יותר, רק נזכיר שמצאנו נקודות חריגות במידע שלנו , נוציא אותם מהחישוב ונבדוק האם אנו מקבלים תוצאות טובות יותר.

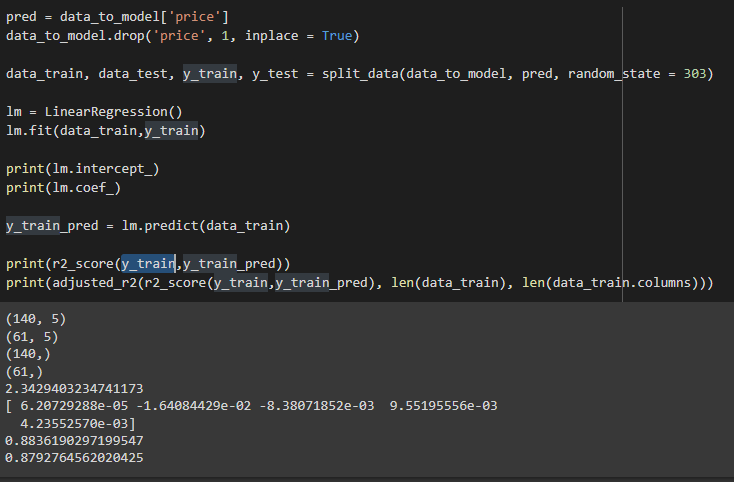


Text

Description automatically generated

ניתן לראות שהמודל עלה באחוזי חיזוי מ 81 ל85 , הנקודות החריגות באמת גרמו לנו לחזות בצורה גרועה יותר.

עכשיו ננסה להוסיף למודל את הפרמטרים עם הקורלציה החזקה למחיר.



גם פה ניתן לראות שיפור משמעותי , 88% אחוז חיזוי.