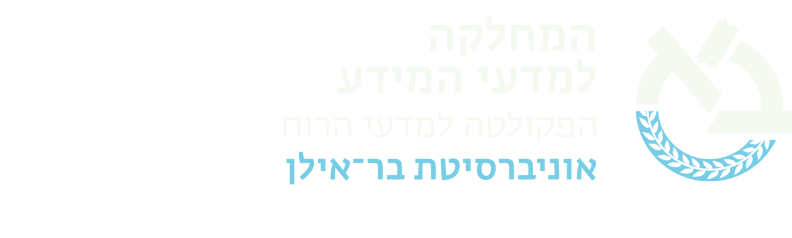
ויזואליזציה של נתונים -

עבודת גמר



|  |  |  |
| --- | --- | --- |
| **מגישים:** |  |  |
|  | במסגרת קורס ״ויזואליזציה של נתונים ״  גב׳ שיר מאמיה |  |

**תיאור סט הנתונים**

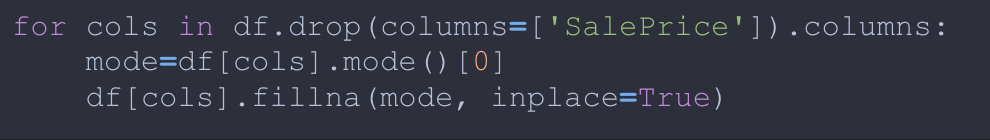
סט הנתונים אותו בחרתי הינו House Prices, נלקח מאתר Kaggle (מצ״ב [קישור](http://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview/description)).

פרטים אודות הנתונים -

1. מדובר בקובץ של נתונים אודות נכסים שנמכרו בעיירה Ames שבמדינת Iowa בארה״ב, כאשר עבור כל נכס מובאים כ - 79 מאפיינים שונים.   
   \*\* לציין כי היות ומדובר בפרויקט ML, הנתונים מכילים שני קבצים (train ו - test), לטובת העבודה השתמשתי רק בקובץ נתוני האימון (train) \*\*
2. מטרת האתגר המקורי באתר Kaggle הינה בניית מודל לחיזוי מחירי בתים בהינתן מאפיינים שונים, כאשר מטרת המשנה הינה מציאת הפיצ׳רים הרלוונטיים ביותר לחיזוי המחיר. כמובן שבעבודה זו לא אמשיך לשלב החיזוי אך כן אתייחס לפיצ׳רים המרכזיים הרלוונטיים.
3. הקובץ מכיל כ - 1460 דוגמאות (rows) וכאמור לעיל, עבור כל דוג׳ ישנן כ - 79 פיצ׳רים + מזהה ייחודי לנכס = 80 עמודות (columns) - הנתונים הובאו בעזרת פקודת shape ב - pandas.
4. מדובר בפיצ׳רים מסוגים שונים, גם נומרים (float64 ו - 64int) וגם קטגוריאליים (object), אך ללא עמודות תאריך - הנתונים נבחנו בעזרת פקודת info.
5. הפיצ׳רים עצמם רבים ומגוונים, ביניהם ניתן למצוא את השכונה בה נמצא הבית, גודל הבית, כמות חדרים, גודל מרתף, מספר חניות, מדד כללי לאיכות הבית וכד׳.
6. בנוסף, וגם כאן באמצעות פקודות info, ניתן לראות כי ישנן עמודות עם ערכים חסרים, כאשר חלקן עם מספר רב של ערכים חסרים.

התמודדות עם ערכים חסרים -

1. היות ול - 6 מהפיצ׳רים ('Alley','MasVnrType', 'FireplaceQu','PoolQC','Fence','MiscFeature') ישנה כמות רבה של ערכים חסרים, החלטתי להיפטר מפיצ׳רים אלה באמצעות פקודת drop.
2. עבור יתר הפיצ׳רים היות והיה מדובר בכמות מועטה של ערכים חסרים החלטתי למלאם באמצעות השכיח (mode) תוך יצירת לולאה פשוטה כפי שניתן לראות כאן -



כעת הדאטה ללא ערכים חסרים, ואנו יכולים לעבור לשלבים הבאים.

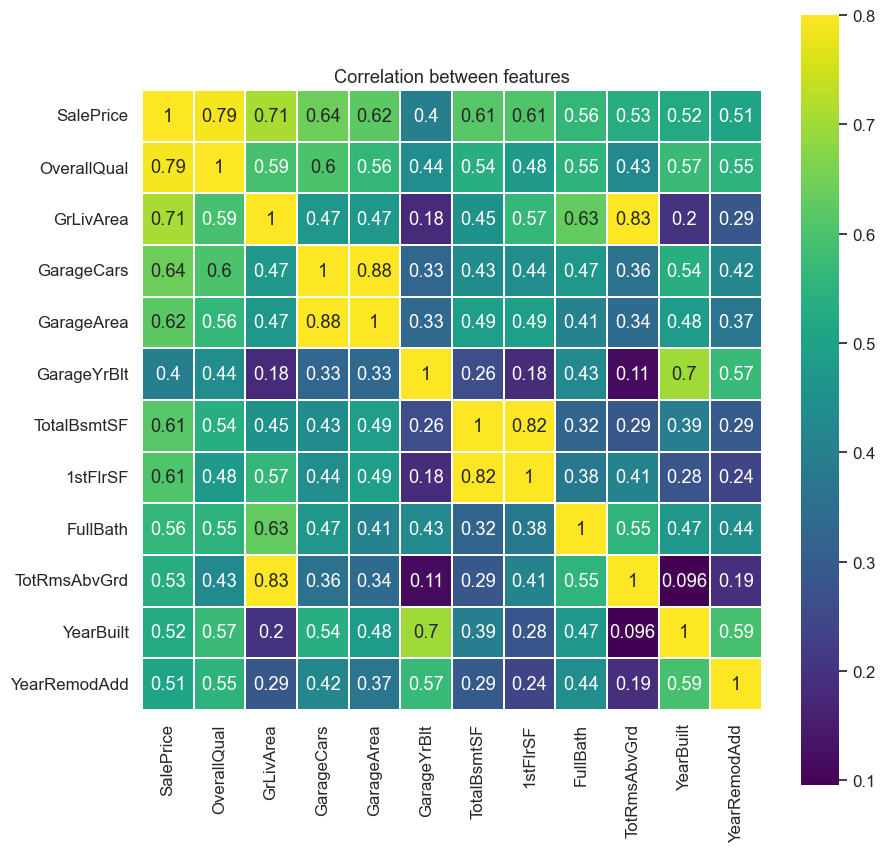
**ניסוח הבעיה, שאלת המחקר, השערות וציפיות** -

1. השאלה המרכזית עימה אתמודד בעבודה הינה ״מהם המשתנים הרלוונטיים ביותר לטובת השפעה על מחיר נכס״, כאשר ניתן לקחת את המסקנות אפילו ״צעד אחד קדימה״ ולבנות על בסיס התוצרים מודל לחיזוי מחיר בית, במקרה כזה השאלה תהיה ״האם ניתן לחזות מחיר בית בהינתן נתונים מסוימים.
2. בהתבסס על היכרות בסיסית יחסית עם עולם תוכן זה, אוכל לשער/לצפות שמשתנים כדוג׳ גודל הבית, מספר חדרים, שכונה בה ממוקם הבית וכד׳ יהיו המשתנים בעלי ההשפעה הרבה ביותר.
3. בהמשך ישיר לכך, יהיה מעניין לראות האם קיימים פיצ׳רים מעולמות תוכן אחרים אשר בהסתכלות ראשונית לא הייתי מצפה שיהיו בעלי השפעה משמעותית על המחיר, או אפילו מספר פיצ׳רים שלכל אחד מהם בפני עצמו השפעה חלשה יחסית, אך השילוב של כולם ביחד מהווה גורם משפיע משמעותי.

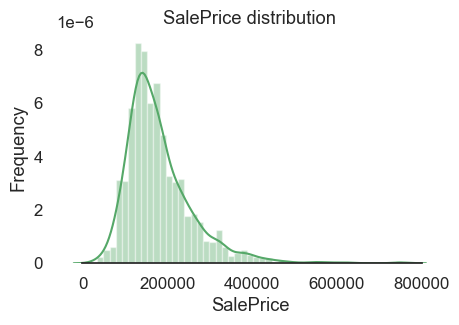
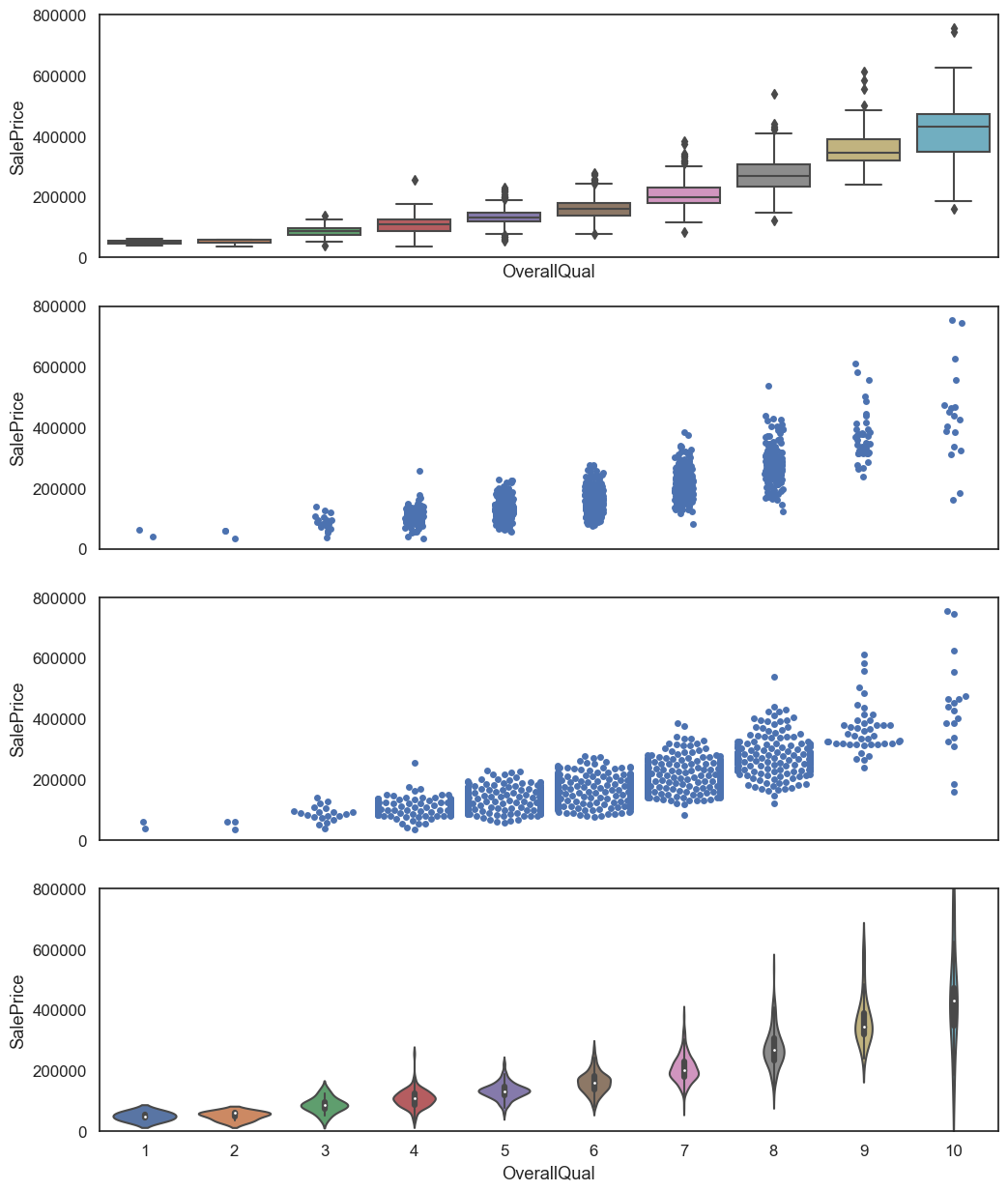
**ניתוח הנתונים (EDA)** -

1. ראשית נריץ את פקודת describe, ע״מ ללמוד קצת יותר על הנתונים (מצ״ב גזיר מהתוצאות) -   
   
2. בעזרת פקודה זו ניתן ללמוד על התפלגות הנתונים בפיצ׳רים השונים, לא אפרט פה בכוונה משום שניגע בכל פיצ׳ר לגופו.

מציאת קורולציה בין משתנים למטרה (מחיר) -

1. ע״מ למצוא מי הפיצ׳רים היותר משמעותיים ולהתמקד בניתוחם, החלטתי בחון את כלל המשתנים ביחס למשתנה המטרה שלי, במקרה הזה SalePrice.
2. ראשית הרצתי את כלל הפיצ׳רים לטובת הצגתם בגרף, כולל התייחסות לעובדה של האם מדובר בקורולציה שלילית/חיובית (משמאל ל - 0 שלילי, ומימין חיובי). מצ״ב הגרף -   
   
3. החלטתי להתמקד בפיצ׳רים אשר השפעתם (בערך מוחלט) הינה מעל 0.5, ורק אותם להציג במטריצת קורולציות -   
     
   העמודה השמאלית ביותר הינה הקורלציה בין המשתנים השונים לפיצ׳ר המטרה (מחיר הבית) לפי הסדר מהגדול לקטן (בעל ההשפעה הרבה ביותר לנמוך ביותר), שוב, לאחר סינון של המשפיעים ב - 0.5 ומעלה.
4. את הפיצ׳רים הנ״ל אחקור כעת באופן מעמיק יותר.

חקר הפיצ׳רים המשמעותיים ביותר -

1. ראשית נסתכל באופן כולל על פיצ׳ר המטרה (מחירי הבתים), ניתן לראות כי מדובר בהתפלגות נורמלית עם ״זנבות״ כצפוי -   
   
2. OveralQual -
   1. פיצ׳ר המתאר את איכות החומרים והגימור בבניית הבית, כפי שניתן לראות מעלה, הינו הפיצ׳ר המוביל בקורלציה למחיר.
   2. להלן הצגת הקשר בין המשתנה הנ״ל למחיר, בארבעה גרפים שונים (boxplot,   
      stripplot,swarmplot,violinplot), כאשר כולם מציגים למעשה אותו דבר אך באופנים שונים -   
      
   3. בכל הגרפים לעיל ניתן לראות שככל שעולה האיכות הכללית (ציר ה - x) כך עולה המחיר (ציר ה - y), אך מכל גרף ניתן להסיק תובנות ויזואליות גם על התפלגות הנתונים.