תרגיל מסכם

חלק א

**לינק למקור הנתונים:** <https://www.kaggle.com/datasets/hm-land-registry/uk-housing-prices-paid>  
  
**נתונים טכניים על המאגר:** 22489348 שורות, 11 עמודות ו2.41 GB סך הנפח המידע.  
  
**תיאור כללי של המאגר:** המאגר שנבחר הוא מאגר הנתונים של מחירי הבתים בבריטניה מ-Kaggle, והוא מכיל מידע על כל המכירות הרשומות של נכסים באנגליה ובוויילס שנמכרו במחיר שוק מלא.   
להלן פירוט על מאפייני הדאטה והמידע הכללי שנמצא במאגר זה:

**מקורות המידע:** המידע במאגר מגיע מרשם המקרקעין הבריטי (HM Land Registry) וכולל נתונים מאז ינואר 1995.

**תכנים במאגר:** המאגר מכיל את השדות הבאים:

**Transaction unique identifier:** מזהה ייחודי לכל עסקה, מספר זה נוצר אוטומטית ומוקצה לכל מכירה.

**Price:** מחיר המכירה כפי שמופיע בשטר ההעברה.

**Date of Transfer:** תאריך השלמת המכירה.

**Property Type:** סוג הנכס (D = Detached, S = Semi-Detached, T = Terraced, F = Flats/Maisonettes, O = Other).

**Old/New:** מציין האם הנכס הוא חדש או מבנה קיים (Y = נכס חדש, N = מבנה קיים).

**Duration:** סוג הבעלות (F = Freehold, L = Leasehold וכו').

**Town/City:** שם העיר או היישוב.

**District:** שם האזור.

**County:** שם המחוז.

**PPD Category Type:** סוג העסקה (A = מכירה רגילה, B = עסקאות נוספות כמו מכירות בעקבות עיקול או מכירות למשכירים).

**Record Status:** סטטוס הרשומה (A = תוספת, C = שינוי, D = מחיקה).

**מה אין במאגר:** פרטי כתובת מלאים, סוגים מסוימים של נכסים כמו בתי קיט ובונגלוס מסוימים, עסקאות של השכרת נכסים לתקופות קצרות (פחות מ-7 שנים).

**מאפיינים טכניים:  
גודל הנתונים:** 22489348 שורות ו11 עמודות.  
**פורמט הנתונים:** קובץ CSV.

**תיאור מפורט של עמודות חשובות:**  
1. **Price (מחיר):  
תיאור:** עמודה זו מציינת את מחיר המכירה של הנכס כפי שמופיע בשטר ההעברה. זהו המחיר המלא שהנכס נמכר עבורו.

**חשיבות:** מחיר המכירה הוא אחד המדדים המרכזיים לניתוח שוק הנדל"ן. הוא מאפשר לזהות מגמות בשוק, להשוות בין סוגי נכסים שונים, לבחון השפעות כלכליות אזוריות ולבצע הערכות שווי נכסים.

**שימושים פוטנציאליים:** ניתוח מגמות מחירים לאורך זמן, השוואת מחירי נכסים לפי סוג הנכס, אזור או תקופה, הערכת שווי נכסים לצורכי מכירה, קנייה או השקעה, זיהוי אזורים עם עליית מחירים גבוהה להשקעות עתידיות.2. **Date of Transfer (תאריך ההעברה):**  
**תיאור:** עמודה זו מציינת את התאריך שבו הושלמה המכירה של הנכס. זהו התאריך כפי שמופיע בשטר ההעברה.

**חשיבות:** תאריך ההעברה מאפשר ניתוח של מגמות לאורך זמן, זיהוי תקופות של גידול או ירידה בשוק הנדל"ן, והשוואה של פעילות השוק לאורך שנים. נתון זה קריטי לבחינת עונתיות בשוק הנדל"ן ולמעקב אחרי שינויים תקופתיים.

**שימושים פוטנציאליים:** ניתוח עונתיות בשוק הנדל"ן (לדוגמה, האם יש יותר מכירות בקיץ או בחורף), זיהוי תקופות של פעילות מוגברת או ירידה במכירות, השוואת מכירות לפי תקופות כלכליות שונות (לפני ואחרי משברים כלכליים), תכנון זמני מכירה אופטימליים עבור מוכרים.

3. **Old/New (ישן/חדש):  
תיאור:** עמודה זו מציינת אם הנכס הוא חדש או מבנה קיים. ערכים אפשריים: Y = נכס חדש, N = מבנה קיים.

**חשיבות:** הבחנה בין נכסים חדשים לקיימים חשובה לניתוח דינמיקת השוק. נכסים חדשים יכולים להעיד על פיתוח ובנייה חדשה באזור, בעוד שנכסים קיימים משקפים את המצב הנוכחי של שוק הנדל"ן באזור.

**שימושים פוטנציאליים:** ניתוח ביקוש לנכסים חדשים לעומת נכסים קיימים, הבנה של השפעת בנייה חדשה על מחירי הנדל"ן באזור מסוים, זיהוי מגמות פיתוח ובנייה באזורים שונים, סיוע בתכנון עירוני והקצאת משאבים לאזורים עם פיתוח מוגבר.

4. **Duration (משך הבעלות):**  
**תיאור:** עמודה זו מציינת את סוג הבעלות על הנכס. ערכים אפשריים כוללים: F = Freehold (בעלות חופשית), L = Leasehold (חכירה) ועוד.

**חשיבות:** סוג הבעלות משפיע על הערך והמשיכה של הנכס. בעלות חופשית נחשבת לרוב ליותר מבוקשת מכיוון שהיא מעניקה לבעל הנכס זכויות מלאות על הנכס. לעומת זאת, חכירה מציינת שהבעלות היא לתקופה מוגבלת ויש תנאים מסוימים על הנכס.

**שימושים פוטנציאליים:** ניתוח ההעדפות של קונים בין נכסים בבעלות חופשית לנכסים בחכירה, הערכת ערך שוק של נכסים על פי סוג הבעלות, זיהוי אזורים עם ריכוז גבוה של נכסים בחכירה למטרות רגולציה ופיתוח מדיניות, סיוע למשקיעים בהבנת התשואה הצפויה מנכסים בעלי סוגי בעלות שונים.

5. **PPD Category Type (סוג קטגוריית המחיר):**  
**תיאור:** עמודה זו מציינת את סוג העסקה. ערכים אפשריים:

**A:** מכירה רגילה של נכס יחיד במחיר מלא.

**B:** עסקאות נוספות הכוללות מכירות בעקבות עיקול, מכירות למשכירים (Buy-to-Let) ועסקאות אחרות.

**חשיבות:** הבחנה בין סוגי עסקאות מאפשרת ניתוח מדויק יותר של השוק. עסקאות רגילות משקפות את המחירים בשוק החופשי, בעוד שעסקאות מסוג B יכולות להצביע על תנאים מיוחדים כמו מכירות עיקול או עסקאות שאינן למחירים מלאים.

**שימושים פוטנציאליים**: ניתוח שוק נפרד לעסקאות רגילות ולמכירות בעקבות עיקול, הבנה של השפעת עסקאות מסוג B על מחירי הנדל"ן הכלליים, זיהוי מגמות עיקול באזור מסוים כמדד כלכלי, בחינת ההשפעה של קניית נכסים למשכירים על שוק השכירות.

6. **Record Status (סטטוס הרשומה):**  
**תיאור:** עמודה זו מציינת את הסטטוס של הרשומה. ערכים אפשריים:

**A:** תוספת חדשה.

**C:** שינוי ברשומה קיימת.

**D:** מחיקה של רשומה.

**חשיבות**: סטטוס הרשומה חשוב לניהול ומעקב אחרי השינויים במאגר הנתונים. הוא מאפשר לדעת אם הרשומה היא חדשה, אם נעשו בה שינויים כלשהם, או אם היא נמחקה מהמאג. מידע זה קריטי לשמירה על שלמות ועדכניות הנתונים.

**שימושים פוטנציאליים:** מעקב אחר שינויים ועדכונים במאגר המידע, זיהוי טעויות או תיקונים שנעשו ברשומות הקיימות,ניהול מחיקות רשומות בצורה מסודרת לשמירה על דיוק המידע,הבטחת עקביות הנתונים לאורך זמן לצורך ניתוחים מדויקים.

**תיאור מערכת סכמתית תיאורטית שאוספת מעכלת ומעבדת את המידע לכדי תורתו הסופית (= תיאור מילולי של capture ingest store compute שיוביל למידע שהורדנו):**המערכת הסכמתית שנתאר כעת מאפשרת איסוף, טעינה, אחסון ועיבוד של נתוני שוק הנדל"ן בצורה יעילה ומסודרת, ומובילה להפקת מידע בעל ערך המאפשר הבנת מגמות שוק, השוואת מחירים והפקת תובנות כלכליות ומסחריות.

**שלב 1: Capture (איסוף נתונים)**

בשלב זה, המערכת תאסוף את הנתונים ממקורות חיצוניים, במקרה זה ממאגר הנתונים של רשם המקרקעין הבריטי (HM Land Registry). האיסוף יתבצע בצורה אוטומטית באמצעות API או הורדות קבועות של קובצי CSV מאתר Kaggle.

**תהליך:** יצירת חיבור ל-API של HM Land Registry או ל-Kaggle, הורדת הנתונים בצורה תקופתית (יומי/שבועי/חודשי), שמירת הנתונים הגולמיים במערכת האחסון.

**שלב 2: Ingest (טעינת נתונים)**

בשלב זה, הנתונים שנאספו מועברים למערכת עיבוד מרכזית. הנתונים הגולמיים מועתקים, נטענים ומותאמים למבנה שמתאים למערכת האחסון והעיבוד.

**תהליך:** טעינת הנתונים מהמקורות הגולמיים למערכת האחסון (Data Lake או Data Warehouse),ביצוע ניקוי ראשוני של הנתונים: הסרת רשומות כפולות, טיפול בערכים חסרים ותיקון טעויות סינטקטיות, התאמת הנתונים למבנה הרצוי, כולל הקצאת מזהים ייחודיים לרשומות אם נחוץ.

**שלב 3: Store (אחסון נתונים)**

בשלב זה, הנתונים המעובדים מאוחסנים במאגרי נתונים ייעודיים לצורך גישה וניתוח בהמשך. המאגר יכול להיות מבוסס על טכנולוגיות כמו SQL Database, NoSQL, או Data Warehouse כמו Amazon Redshift, Google BigQuery וכו'.

**תהליך:** שמירת הנתונים הנקיים והמעובדים במבנה טבלאי או מבנה אחר שמתאים לניתוח, יצירת אינדקסים על מנת לאפשר שאילתות מהירות ויעילות, שמירת גרסאות שונות של הנתונים במידת הצורך (לדוגמה: נתונים מקוריים, נתונים לאחר ניקוי, נתונים לאחר עיבוד).

**שלב 4: Compute (עיבוד נתונים)**

בשלב זה, הנתונים המאוחסנים משמשים לביצוע עיבודים ואנליזות מתקדמות על מנת להפיק תובנות ומידע נוסף.

**תהליך:** ביצוע שאילתות SQL מורכבות לניתוח הנתונים, שימוש בכלי BI (Business Intelligence) כמו Tableau, Power BI או Looker ליצירת דוחות וויזואליזציות, שימוש באלגוריתמי Machine Learning ו-Data Mining להפקת תובנות מתקדמות.

**תוצר סופי: מידע מעובד ומוכן לשימוש**

הנתונים המועבדים כוללים את כל המידע שנדרש לצורך ביצוע ניתוחים שונים על שוק הנדל"ן, כולל מחיר הנכסים, תאריכי העברה, סוגי נכסים, סוגי בעלות, סוגי עסקאות וסטטוס הרשומות. המידע מוצג בצורה שתאפשר למשתמשים לנתח את השוק, לזהות מגמות, ולהפיק תובנות כלכליות ומסחריות.

**דוגמה לזרימת הנתונים:**

**איסוף נתונים (Capture):** הורדת קובצי CSV מ-Kaggle על בסיס חודשי.

**טעינת נתונים (Ingest):** העלאת קובצי ה-CSV ל-Data Lake, ניקוי ראשוני והכנה.

**אחסון נתונים (Store):** שמירת הנתונים הנקיים ב-Data Warehouse, יצירת אינדקסים וניהול גרסאות.

**עיבוד נתונים (Compute):** ביצוע שאילתות לניתוח, יצירת דוחות וויזואליזציות, הפקת תובנות באמצעות Machine Learning.

**תיאור מילולי וגם הפקודות שהרצנו כדי להעלות את הדאטה הגדול לתוך הביג דאטה Warehouse שבחרנו**

בחרנו להשתמש ב-Warehouse של DuckDB לאחסון ועיבוד נתונים. DuckDB הוא מנוע מסד נתונים קל משקל, יעיל ומיועד לפעולה בסביבות נתונים גדולים. הוא מתאים במיוחד לטעינה מהירה של קבצי CSV ולניתוח עמוק של הנתונים.

**הסבר על הפקודות שהורצו:**

**הגדרת שם לקובץ CSV:**csv\_file\_name = 'price\_paid\_records.csv'  
בפקודה זו, אנו מגדירים את השם של קובץ ה-CSV שבו מצוינים הנתונים שיש לטעון ל-Warehouse.  
  
**חיבור ל-Warehouse של DuckDB:**  
con = duckdb.connect()  
בשורה זו אנו מתחברים ל-Warehouse של DuckDB באמצעות החיבור המוגדר במשתנה 'con'.

**יצירת טבלה וטעינת הנתונים מהקובץ CSV:**  
con.execute("""  
 CREATE TABLE price\_paid\_records AS   
 SELECT \* FROM read\_csv\_auto(?,   
 types={  
 'Transaction unique identifier': 'VARCHAR',  
 'Price': 'INTEGER',  
 'Date of Transfer': 'DATE',  
 'Property Type': 'VARCHAR',  
 'Old/New': 'VARCHAR',  
 'Duration': 'VARCHAR',  
 'Town/City': 'VARCHAR',  
 'District': 'VARCHAR',  
 'County': 'VARCHAR',  
 'PPDCategory Type': 'VARCHAR',  
 'Record Status - monthly file only': 'VARCHAR'  
 }  
 )  
 """, [csv\_file\_name])

בפקודה זו אנו מבצעים שני פעולות:

משתמשים בפונקציה **'read\_csv\_auto'** שמאפשרת טעינה אוטומטית של קובץ CSV לתוך טבלה ב-DuckDB.

מציינים את שמות העמודות ואת סוגי הנתונים שלהם במאגר הנתונים (VARCHAR, INTEGER, DATE וכו') על מנת להגדיר את מבנה הטבלה.

באמצעות פקודות אלו, אנו מבצעים את הטעינה הראשונית של הנתונים מהקובץ CSV אל תוך מאגר הנתונים של DuckDB, ומאפשרים לנו להתחיל בעיבוד ובנית דוחות על הנתונים במערכת.

**השאלות ששאלנו על הדטאה, מה אנחנו יכולים ללמוד מהם ואלו תובנות אנחנו רואים**

**השאלות ששאלנו על הדטאה הן:**1. באילו חודשים מתבצעות הכי הרבה עסקאות?  
2. מה ההבדל במחיר הממוצע בין נכס חדש לנכס ישן עבור כל סוג נכס?  
3. איך משתנה המחיר הממוצע לנכס לאורך השנים?  
4. אילו תקופות התאפיינו בעלייה או ירידה חדה במחירי הנדל"ן?  
5. באילו ערים הייתה העלייה הגדולה ביותר במחיר הממוצע לנכס בין שתי תקופות זמן שונות?  
6. מהם מחירי הנכסים הממוצעים לפי סוג הנכס במחוזות או בערים השונות?  
  
**מה אנחנו יכולים ללמוד מהשאלות:**

1. לזהות עונתיות ודפוסים חודשיים בפעילות העסקאות בשוק הנדל"ן.  
2. להבין את ההבדלים במחירים בין נכסים חדשים וישנים לפי סוג הנכס, מה שעשוי לסייע בקבלת החלטות רכישה.  
3. לעקוב אחר השינויים במחירי הנדל"ן לאורך זמן ולזהות מגמות עולות או יורדות.  
4. לאתר תקופות של תנודות חריגות במחירים, שעשויות להצביע על אירועים או גורמים כלכליים משמעותיים.  
5. לזהות ערים או אזורים ספציפיים שחוו גידול משמעותי במחירי הנדל"ן, מה שעשוי להעיד על הזדמנויות השקעה.  
6. להשוות מחירי נדל"ן ממוצעים לפי סוג הנכס ולפי מיקום גיאוגרפי, מה שעשוי לסייע בהערכת ערכי נכסים ובהחלטות מכירה/רכישה.

**התובנות שאנחנו רואים:**

1. ייתכן שנראה דפוסים עונתיים בפעילות העסקאות, עם עלייה בחודשים מסוימים ושפל בחודשים אחרים.  
2. ייתכן שנראה הבדלים משמעותיים במחירים הממוצעים בין נכסים חדשים לישנים, תלוי בסוג הנכס.  
3. נוכל לזהות מגמות עליות או ירידות במחירי הנדל"ן לאורך השנים, ולקשר אותן לתנאים כלכליים או גורמים אחרים.  
4. נוכל לזהות תקופות ספציפיות עם תנודות חריגות במחירים, ולנסות להבין את הגורמים שהובילו לכך.  
5. נוכל לזהות ערים או אזורים מסוימים שהיו "נקודות חמות" עם עליות חדות במחירי הנדל"ן, מה שעשוי להצביע על הזדמנויות השקעה עתידיות באזורים אלה.  
6. נוכל להשוות מחירים ממוצעים לפי סוג הנכס ולפי מיקום גיאוגרפי, ולזהות אזורים או סוגי נכסים שמציעים ערך טוב יותר או פחות.

השאילתות שעונות על השאלות והשאילתה שבוחרת שורות לדוגמא מכל טבלה

**1. השאילתה שעונה על השאלה:**transactions\_by\_month = duckdb\_to\_pandas('''  
 SELECT DISTINCT  
 strftime('%m', "Date of Transfer") AS Month,  
 COUNT(\*) OVER (PARTITION BY strftime('%m', "Date of Transfer")) AS Transactions  
 FROM price\_paid\_records  
 ORDER BY Month ASC  
''')

**השאילתה שבוחרת שורות לדוגמא:**SELECT \* FROM transactions\_by\_month LIMIT 50 **2. השאילתה שעונה על השאלה:**avg\_price\_by\_type\_status = duckdb\_to\_pandas('''  
 SELECT DISTINCT  
 "Property Type",   
 "Old/New",   
 AVG(Price) OVER (PARTITION BY "Property Type", "Old/New") AS AvgPrice   
 FROM price\_paid\_records  
 ORDER BY "Property Type", "Old/New"  
''')

**השאילתה שבוחרת שורות לדוגמא:**SELECT \* FROM avg\_price\_by\_type\_status LIMIT 50  
  
**3. השאילתה שעונה על השאלה:**avg\_price\_by\_year = duckdb\_to\_pandas('''  
 SELECT DISTINCT  
 strftime('%Y', "Date of Transfer") AS Year,   
 AVG(Price) OVER (PARTITION BY strftime('%Y', "Date of Transfer")) AS AvgPrice   
 FROM price\_paid\_records  
 ORDER BY Year ASC  
''')

**השאילתה שבוחרת שורות לדוגמא:**SELECT \* FROM avg\_price\_by\_year LIMIT 50  
  
**4. השאילתה שעונה על השאלה:**top\_cities\_price\_increase = duckdb\_to\_pandas('''  
 WITH avg\_price\_per\_city AS (  
 SELECT DISTINCT "Town/City", strftime('%Y', "Date of Transfer") AS Year, AVG(Price) OVER (PARTITION BY "Town/City", strftime('%Y', "Date of Transfer")) AS AvgPrice  
 FROM price\_paid\_records  
 )  
 SELECT a."Town/City",   
 a.AvgPrice AS AvgPrice\_1995,   
 b.AvgPrice AS AvgPrice\_2017,   
 (b.AvgPrice - a.AvgPrice) AS PriceIncrease  
 FROM avg\_price\_per\_city a  
 JOIN avg\_price\_per\_city b ON a."Town/City" = b."Town/City"  
 WHERE a.Year = '1995' AND b.Year = '2017'  
 ORDER BY PriceIncrease DESC  
 LIMIT 10  
''')

**השאילתה שבוחרת שורות לדוגמא:**SELECT \* FROM top\_cities\_price\_increase LIMIT 50

**5. השאילתה שעונה על השאלה:**price\_changes = duckdb\_to\_pandas('''  
 WITH monthly\_prices AS (  
 SELECT   
 strftime('%Y-%m', "Date of Transfer") AS YearMonth,   
 AVG(Price) AS AvgPrice  
 FROM price\_paid\_records  
 GROUP BY YearMonth  
 ),  
 price\_changes AS (  
 SELECT   
 YearMonth,   
 AvgPrice,   
 LAG(AvgPrice) OVER (ORDER BY YearMonth) AS PrevAvgPrice  
 FROM monthly\_prices  
 )  
 SELECT   
 YearMonth,   
 AvgPrice,   
 PrevAvgPrice,   
 (AvgPrice - PrevAvgPrice) AS PriceChange,  
 (AvgPrice - PrevAvgPrice) / PrevAvgPrice \* 100 AS PriceChangePercent  
 FROM price\_changes  
 WHERE PrevAvgPrice IS NOT NULL  
 ORDER BY ABS(PriceChangePercent) DESC  
 LIMIT 10  
''')

**השאילתה שבוחרת שורות לדוגמא:**SELECT \* FROM price\_changes LIMIT 50

**6. השאילתה שעונה על השאלה:**avg\_price\_by\_type\_county = duckdb\_to\_pandas('''  
 SELECT DISTINCT  
 "County",   
 "Property Type",   
 AVG(Price) OVER (PARTITION BY "County", "Property Type") AS AvgPrice   
 FROM price\_paid\_records  
 ORDER BY "County" ASC  
''')

**השאילתה שבוחרת שורות לדוגמא:**SELECT \* FROM avg\_price\_by\_type\_county LIMIT 50

**הסיפור שנרצה לספר – סיפור היסטוריית הנדל"ן באנגליה וויילס (1995-2017)**

מבוא

**שוק הנדל"ן באנגליה וויילס עבר תהפוכות רבות בין השנים 1995 ל-2017. באמצעות הנתונים שנאספו ממאגר מחירי הדיור, נוכל לבחון את המגמות המרכזיות, השינויים במחירים והשפעות גורמים שונים על שוק הנדל"ן.**

השנים המוקדמות (1995-2000)

**בתחילת התקופה, מחירי הנדל"ן היו יציבים יחסית. במהלך שנים אלו:**

**הכלכלה הבריטית התאוששה ממיתון ראשית שנות ה-90.**

**מחירי הדיור עלו באיטיות, עם עלייה מתונה של כ-2% לשנה.**

**הביקוש לנכסים היה יציב, והשוק לא חווה תנודות משמעותיות.**

הבועה והעלייה במחירים (2001-2007)

**השנים האלו התאפיינו בעלייה חדה במחירי הנדל"ן:**

**הבנקים סיפקו משכנתאות בריבית נמוכה, מה שהגביר את הביקוש לנכסים.**

**מחירי הדיור עלו בשיעור שנתי ממוצע של כ-10%.**

**הביקוש לנכסים במרכזי הערים עלה, והביקוש לנכסים בפרברים ירד.**

**המשקיעים ראו בנדל"ן אפיק השקעה משתלם, מה שהגביר את מספר העסקאות.**

המשבר הכלכלי והירידה במחירים (2008-2009)

**המשבר הפיננסי העולמי ב-2008 השפיע בצורה דרמטית על שוק הנדל"ן:**

**מחירי הנדל"ן צנחו בממוצע ב-15% בין 2008 ל-2009.**

**מספר העסקאות ירד בצורה חדה, והביקוש לנכסים ירד.**

**הבנקים הקשיחו את תנאי המשכנתאות, מה שהקשה על קונים פוטנציאליים.**

ההתאוששות והיציבות (2010-2017)

**לאחר המשבר, שוק הנדל"ן החל להתאושש בהדרגה:**

**בין 2010 ל-2013, המחירים עלו במתינות של כ-3% לשנה.**

**משנת 2014 ועד 2017, חלה עלייה מהירה יותר במחירים, במיוחד בערים הגדולות כמו לונדון.**

**השוק התייצב והפך למאוזן יותר, עם עליות מחירים מתונות ויציבות יחסית במספר העסקאות.**

ניתוח לפי סוגי נכסים ומיקומים

סוגי נכסים:

**בתים מנותקים (**Detached**) היו הנכסים היקרים ביותר, עם עליות מחירים חדות במהלך התקופה.**

**דירות (**Flats**) חוו עליות מחירים מתונות יותר, אך הפכו לפופולריות במיוחד בערים הגדולות.**

מיקומים:

**לונדון וסביבתה ראו עליות מחירים חדות ביותר, במיוחד לאחר המשבר הכלכלי.**

**באזורים הכפריים, המחירים נותרו יציבים יותר במהלך התקופה כולה.**

**השיטות שנבחרו "להקטנת" הדאטה**

**Group By & Aggregate Functions:**

**transactions\_by\_month:**

למה נבחרה השיטה: **קיבוץ העסקאות לפי חודש באמצעות GROUP BY ופונקציית COUNT מאפשר לנו להבין את תדירות העסקאות בחלוקה חודשית. זה עוזר לזהות מגמות עונתיות ולבחון את הפעילות בשוק לאורך השנה.**

**avg\_price\_by\_type\_status:**

למה נבחרה השיטה: **שימוש ב-GROUP BY ובפונקציית AVG לחישוב ממוצע מחירים לפי סוג נכס וסטטוס חדש/ישן מאפשר להבין את השוני במחירי הנכסים בהתאם לסוג הנכס ולמצבו. זה מסייע בזיהוי מגמות בשוק ובחינת ההבדלים במחירים בין נכסים חדשים לישנים.**

**avg\_price\_by\_year:**

למה נבחרה השיטה: **קיבוץ הנתונים לפי שנה באמצעות GROUP BY ופונקציית AVG מאפשר לבחון את המגמות השנתיות במחירי הנכסים. זה עוזר לזהות עליות או ירידות מחירים לאורך השנים ולבצע ניתוח ארוך טווח.**

**price\_changes:**

למה נבחרה השיטה: **קיבוץ לפי חודש-שנה לצורך חישוב שינויי מחירים מאפשר לעקוב אחר התפתחות המחירים בטווחי זמן קצרים יותר ולהבין את הדינמיקה של השינויים בשוק הנדל"ן.**

**Window Functions:**

**avg\_price\_by\_type\_status, avg\_price\_by\_year, top\_cities\_price\_increase, avg\_price\_by\_type\_county:**

למה נבחרה השיטה: **פונקציות חלון כמו PARTITION BY ו-OVER מאפשרות לחשב ממוצעים על קבוצות נתונים מבלי להשתמש ב-GROUP BY, מה שמאפשר להוסיף עמודות מחושבות תוך שמירה על מבנה הנתונים המקורי. זה מועיל במיוחד כאשר רוצים לבצע ניתוחים מורכבים ולשמור על גמישות גבוהה בניתוח הנתונים.**

**Filtering:**

**top\_cities\_price\_increase:**

למה נבחרה השיטה: **סינון הנתונים לשנים 1995 ו-2017 מאפשר להתמקד בתקופות זמן ספציפיות בעלות חשיבות אנליטית גבוהה. הגבלת השורות ל-10 הערים עם העלייה הגבוהה ביותר במחירים עוזרת להתמקד במקרים הבולטים ביותר ולזהות מגמות משמעותיות.**

**price\_changes:**

למה נבחרה השיטה: **סינון הנתונים להצגת 10 התקופות עם השינויים החדים ביותר במחירים מאפשר להתמקד באירועים הקריטיים ביותר בשוק הנדל"ן. זה עוזר להבין את התנודתיות בשוק ולזהות גורמים אפשריים לשינויים אלו.**

**Temporal Filtering:**

**top\_cities\_price\_increase ו-price\_changes:**

למה נבחרה השיטה: **התמקדות בטווחי זמן ספציפיים מאפשרת לבחון תקופות או שנים מסוימות שבהן חלו שינויים משמעותיים בשוק הנדל"ן. זה עוזר לזהות מגמות לאורך זמן ולבצע השוואות בין תקופות שונות, מה שיכול לספק תובנות חשובות על השוק.**

**וויזואליזציות נבחרות והסבר על בחירתן**

גרפי עמודות (**Bar Charts**):

עסקאות לפי חודש

תקופות עם שינויים חדים במחירים

**גרפי עמודות נבחרו כדי להמחיש השוואה ויזואלית של ערכים כמותיים בין קטגוריות שונות (כמו חודשים או תקופות). גרפים אלה מאפשרים לזהות בקלות הבדלים בין קטגוריות שונות וערכים חריגים.**

גרף עמודות מרובדות (**Clustered Bar Chart**):

ממוצע מחיר לפי סוג נכס וסטטוס חדש/ישן

**גרפי עמודות מרובדות משמשים להשוואת ערכים בין קבוצות שונות בתוך אותה קטגוריה ראשית (כמו סוג נכס). זה מסייע בזיהוי הבדלים במחירים בין נכסים חדשים לישנים באותו סוג נכס.**

גרף קווי (**Line** **Plot**):

ממוצע מחיר לפי שנה

**גרף קווי נבחר כדי להדגיש את המגמה והשינויים לאורך הזמן (שנים שונות). באמצעות קו ניתן לזהות עליות, ירידות ונקודות תפנית בקלות.**

מה ניתן לראות בגרפים:

בעמודות לפי חודש:

**הגרף מציג את החודשים עם נפח עסקאות גבוה, מה שמצביע על עונתיות ותקופות שיא בשוק הנדל"ן.**

בעמודות המרובדות:

**הגרף מציג בבירור את פערי המחירים בין נכסים חדשים לישנים בכל סוג נכס, ומאפשר לזהות קטגוריות עם פערים גדולים או קטנים.**

בגרף הקווי:

**ניתן לראות את המגמות ארוכות הטווח במחירי הנדל"ן, כגון תקופות של עלייה מתמשכת, ירידות או שינויים חדים לאורך השנים.**

בעמודות של שינויים חדים:

**הגרף מציג תקופות עם שינויים חריגים ביותר במחירי הנדל"ן, כלפי מעלה או מטה, ומסייע בזיהוי תקופות קריטיות בשוק.**

הסבר כללי על הבחירה בוויזואליזציות:

**הוויזואליזציות נבחרו כדי לספק תמונה ברורה ומקיפה של המגמות והדפוסים המרכזיים בנתוני מחירי הנדל"ן, ולאפשר זיהוי קל של תקופות, סוגי נכסים ואזורים בעלי עניין מיוחד לצורך קבלת החלטות ועריכת ניתוחים נוספים. כל גרף נבחר בקפידה כדי להציג את המידע בצורה הברורה והמדויקת ביותר, ולסייע בזיהוי מגמות ותובנות חשובות בשוק הנדל"ן.**

**הסברים איך להריץ**

**1.** התקנת חבילות פייתון נדרשות:

**פתח את**Terminal **וודא שאתה נמצא באותו ספריית העבודה של הקוד. הרץ את הפקודה הבאה כדי להתקין את כל החבילות הנדרשות מתוך הקובץ** requirements.txt**:**

**pip install -r requirements.txt**

**2**. הורדת מאגר הנתונים:

**הקוד עובד עם מאגר נתונים שנקרא "**price\_paid\_records.csv**". אם עדיין לא הורדת את המאגר, ניתן למצוא אותו בקישור הבא:**

https://www.kaggle.com/datasets/hm-land-registry/uk-housing-prices-paid

**הורד את הקובץ** CSV **ושמור אותו באותה ספרייה עם קבצי הקוד.**

**3.** הרצת קובץ ה**-Python** להעלאת הנתונים:

**הרץ את הקובץ** duckdb&sqlite.py **באמצעות הפקודה:**

**python duckdb\_and\_sqlite.py**

**הקובץ הזה יעלה את הנתונים מה-**CSV **לתוך** DuckDB**, יריץ כמה שאילתות ניתוח, ויכתוב את התוצאות לתוך בסיס נתונים** SQLite **חדש בשם "**price\_paid\_records.db**".**

**4.** הרצת הדאשבורד:

**לאחר שהשלב הקודם הסתיים בהצלחה, הרץ את הקובץ** price\_paid\_dashboard.py **כדי להפעיל את הדאשבורד:**

**streamlit run price\_paid\_dashboard.py**

**זה יפתח את הדאשבורד ב-**Streamlit **בדפדפן שלך. תוכל לנווט בין הדפים השונים באמצעות התפריט הצדדי.**

**5.** ניווט בדאשבורד:

**הדאשבורד מכיל כמה דפים:**

**- Questions: מציג את השאלות המרכזיות שהניתוח מנסה לענות עליהן.**

**- Story: מספק סקירה כללית של הסיפור והתובנות העיקריות מהניתוח.**

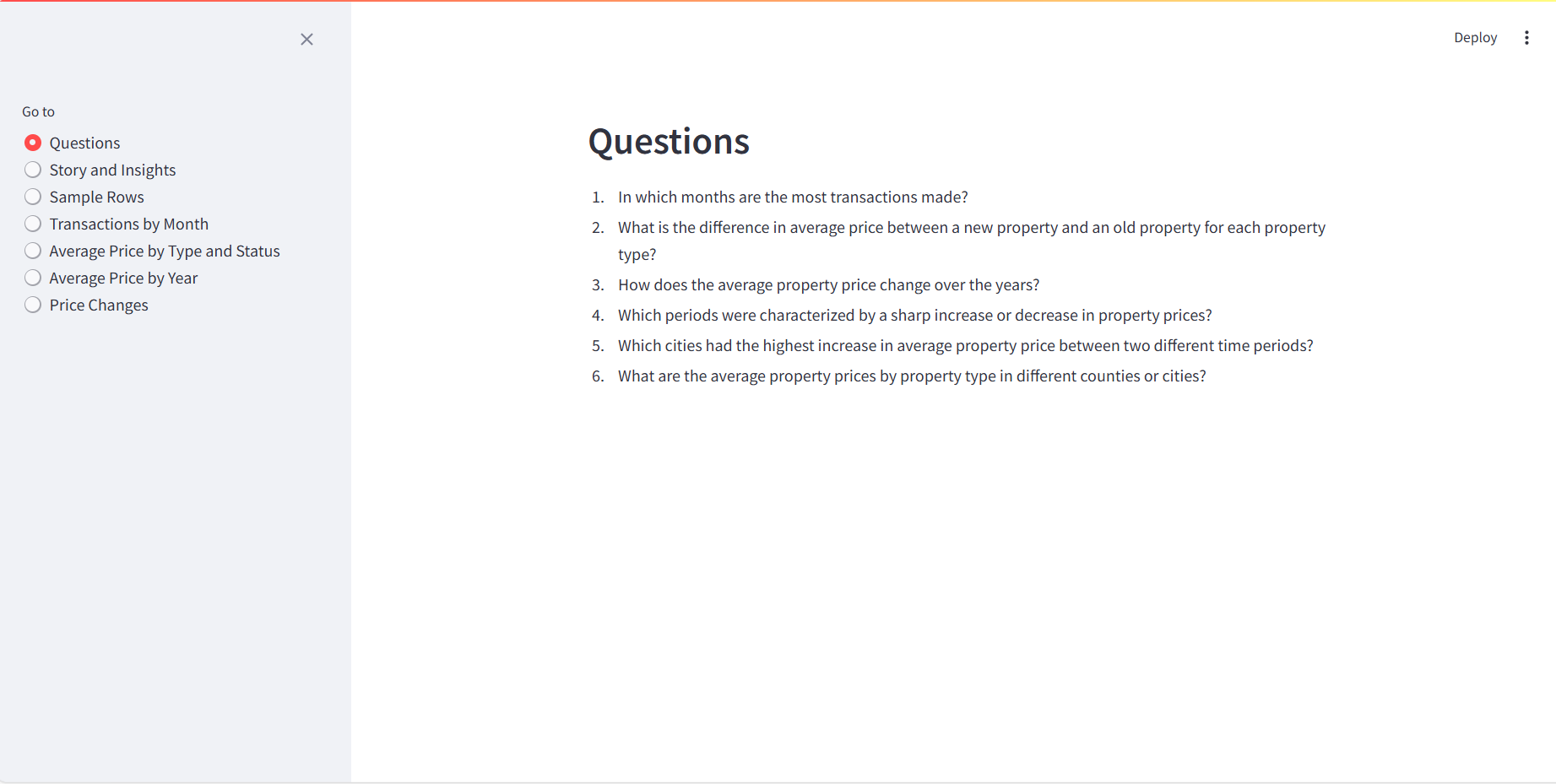
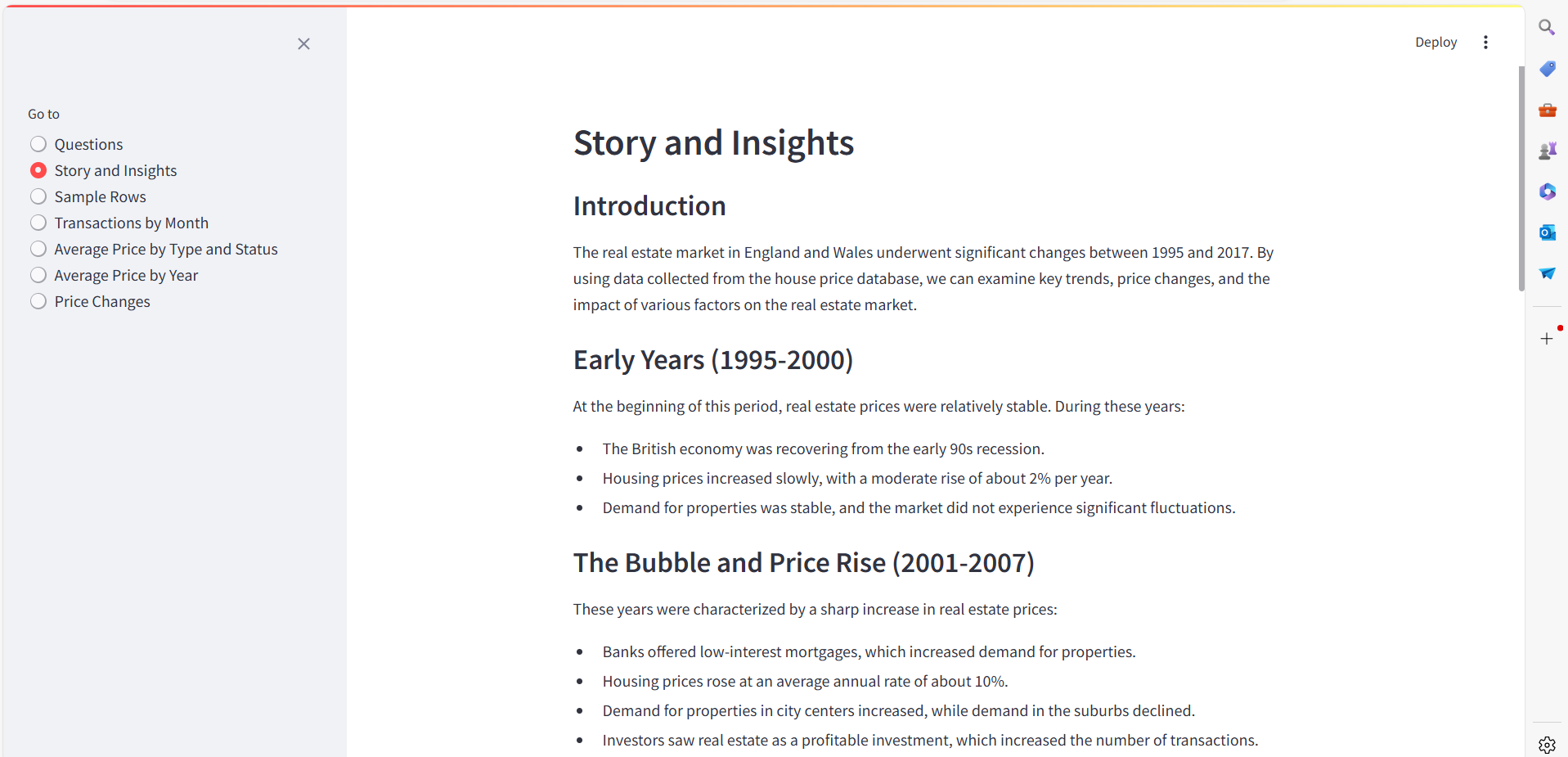
**- Sample Rows: מציג שורות דוגמה מכל הטבלאות שנוצרו.**

**- שאר הדפים מציגים וויזואליזציות שונות ותובנות ממאגר הנתונים.**

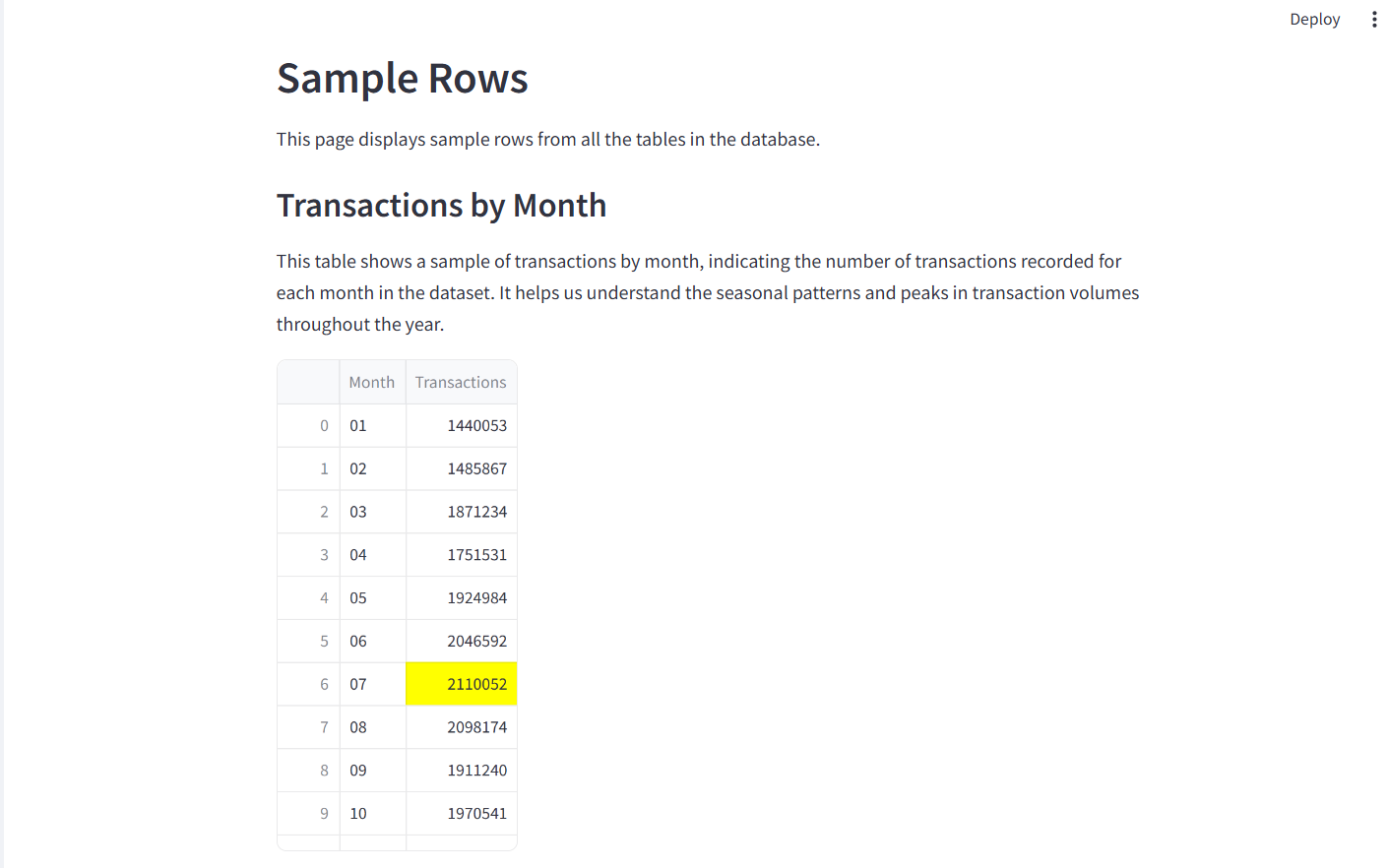
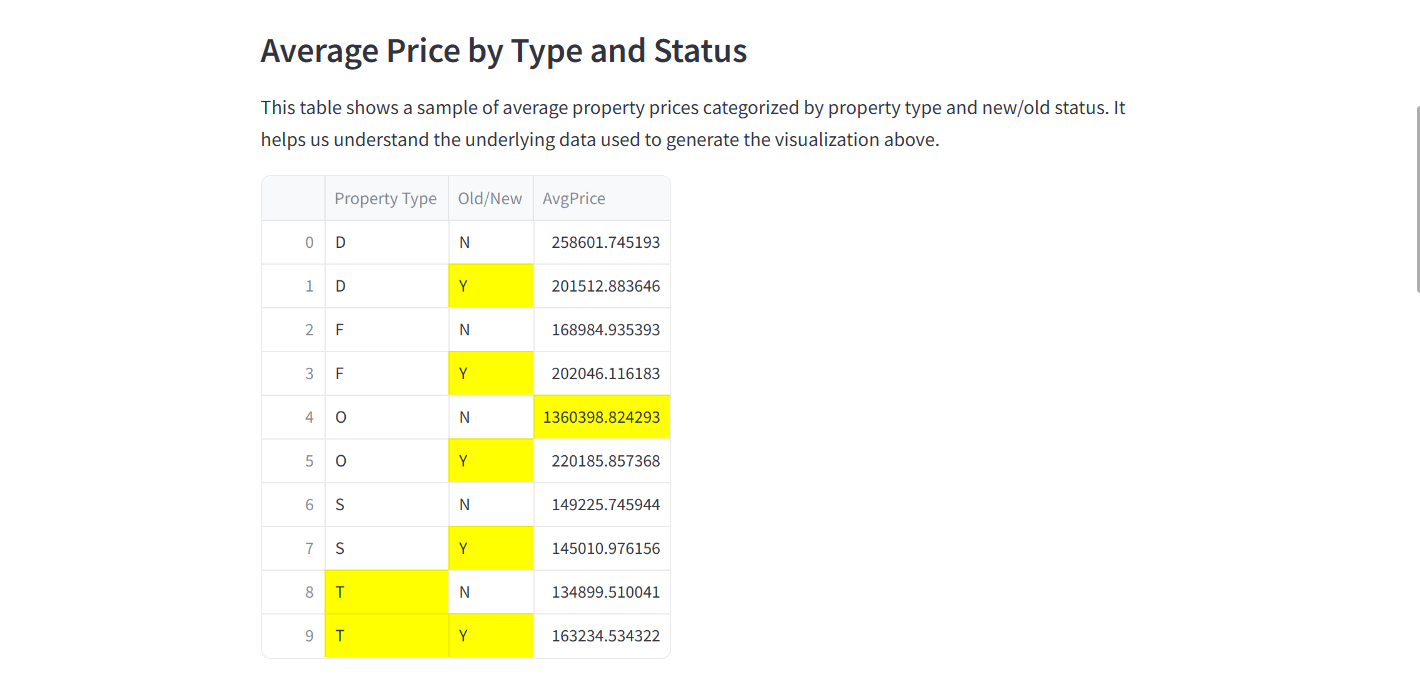
**השתמש בתפריט הצדדי כדי לנווט בין הדפים השונים ולצפות בתוכן.**

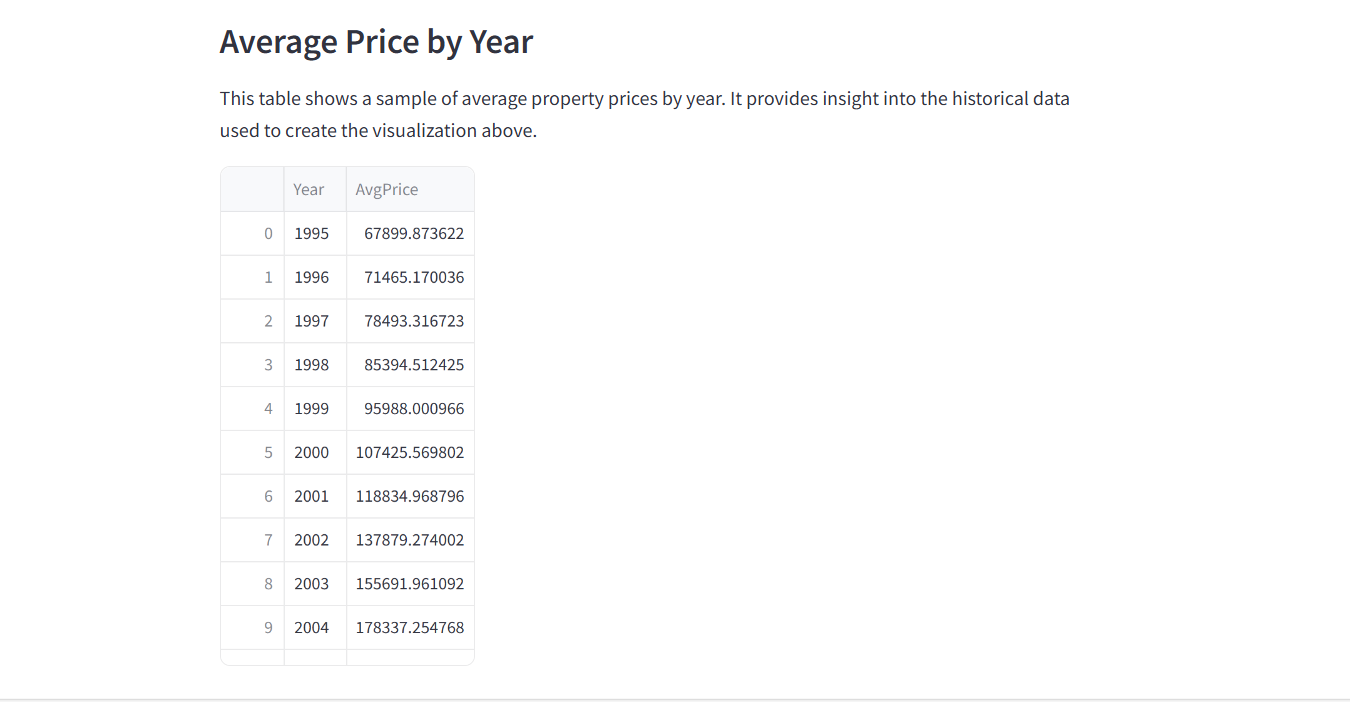
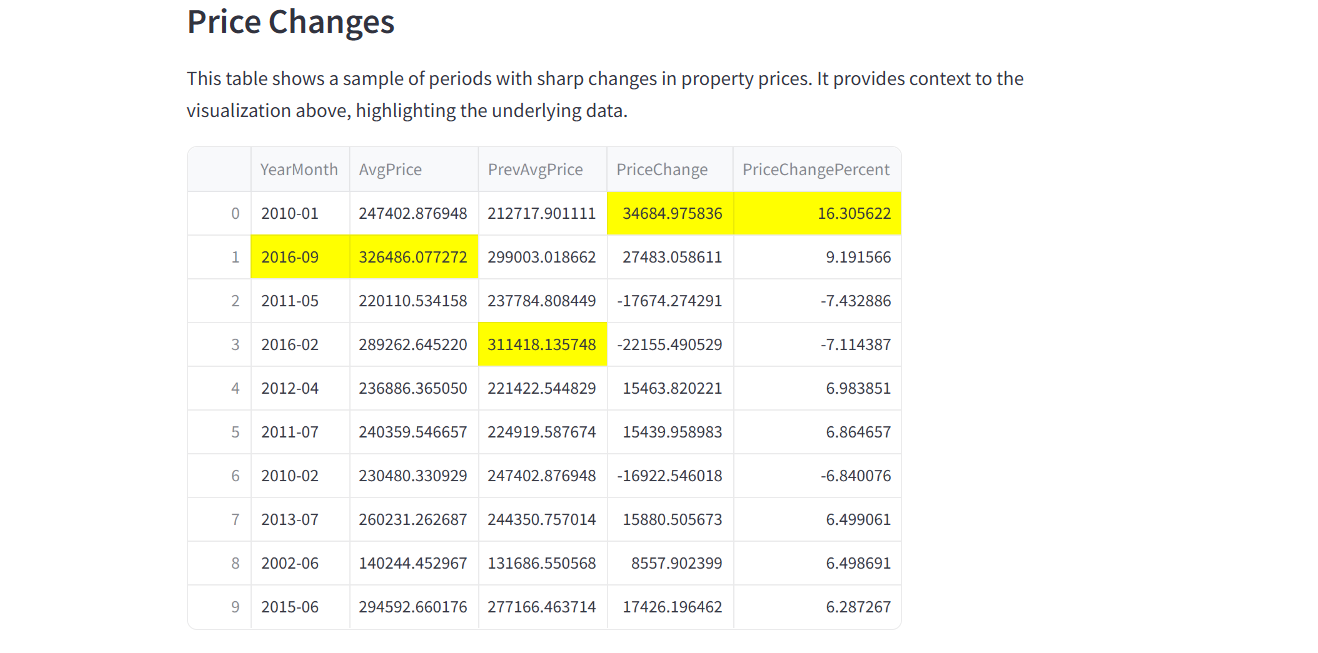
6. **עצירת הדאשבורד:**

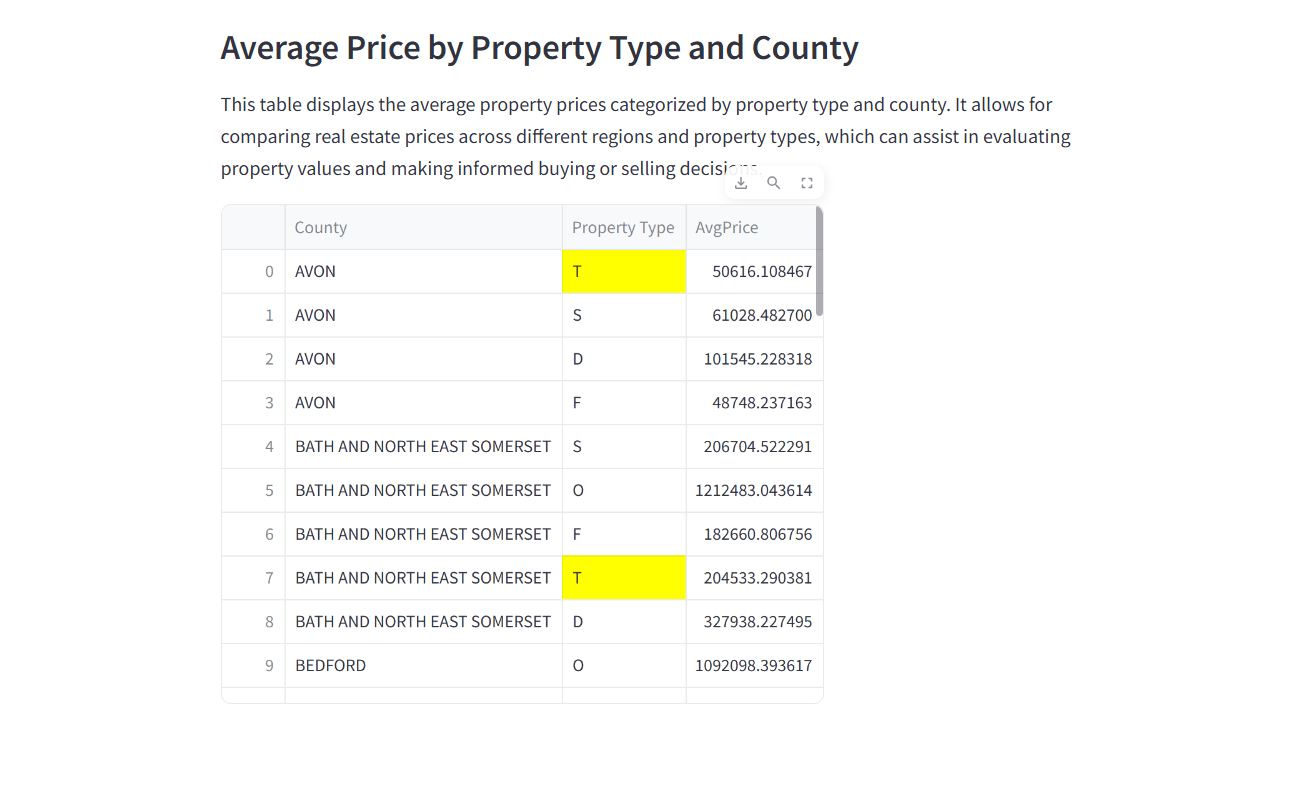
כדי לעצור את הדאשבורד, השתמש בקיצור המקלדת "Ctrl + C" בחלון ה-Terminal שבו הופעל הפקודה של הרצת הדאשבורד.

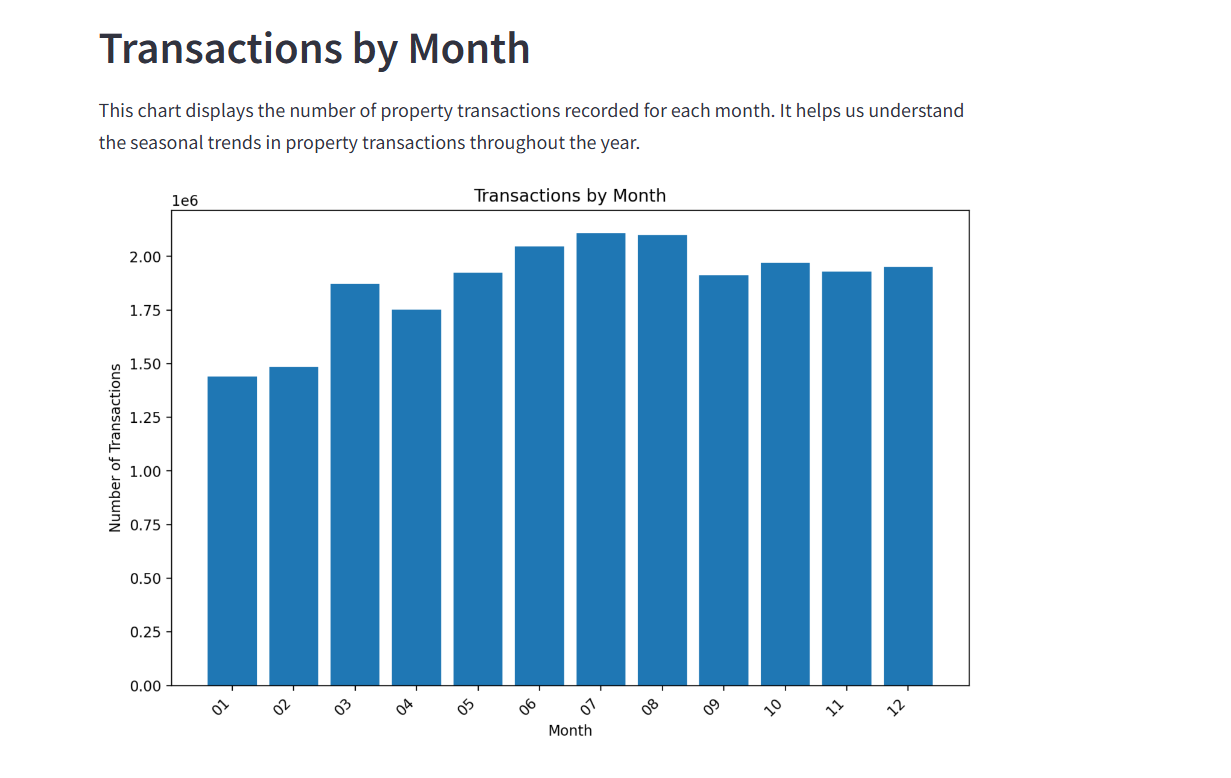
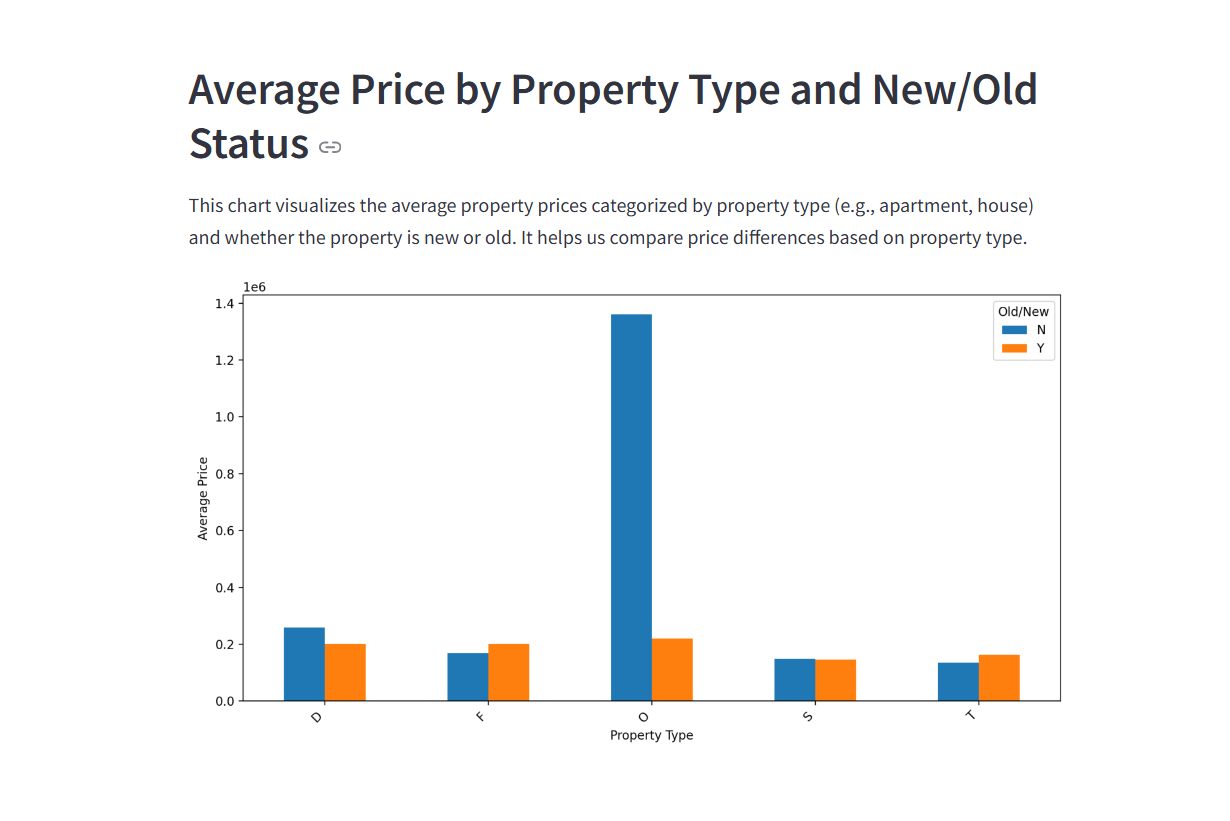
צילומי מסך של הדפים העיקריים בדאשבורד

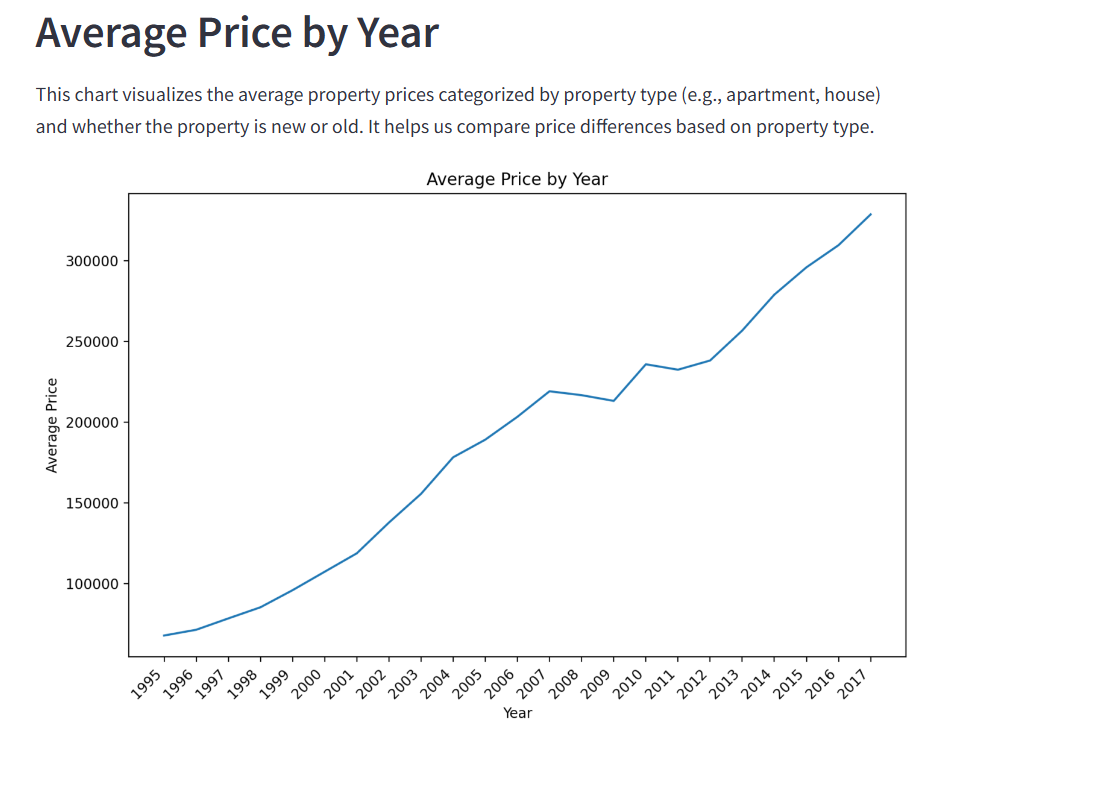
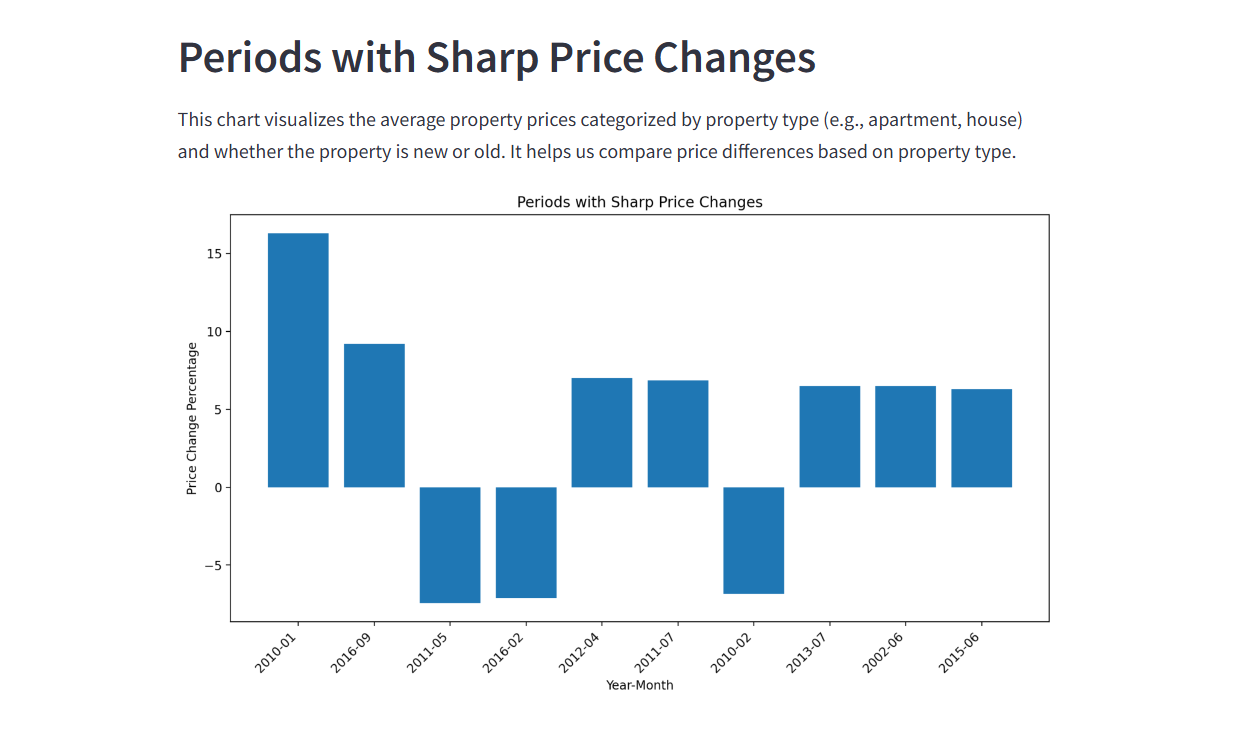


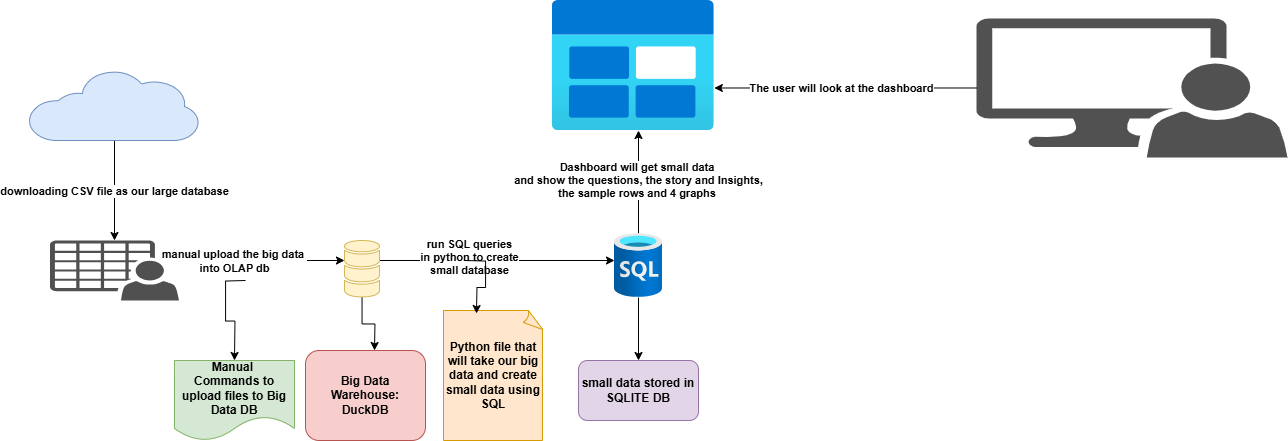








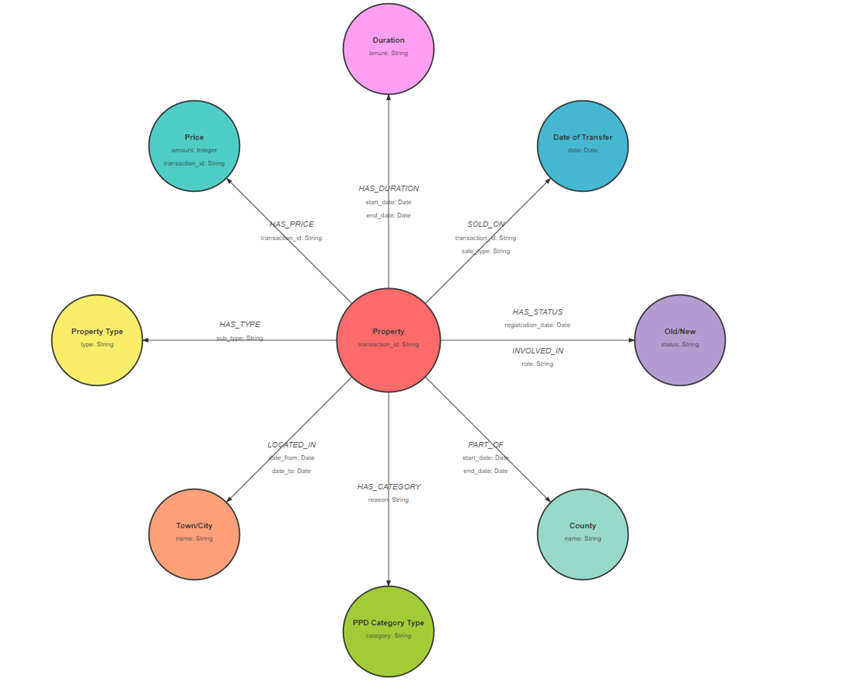


ציור סכמתי של המערכת

האחריות והעבודה של כל אחד מהסטודנטים במהלך התרגיל

**שנינו היינו אחראיים לכל השלבים במהלך התרגיל ולכתיבת הקוד והמסמך**

**חלק ג**

ציור של ה**NODES** וה **RELATIONS**

הסבר לגבי הסכמה: **הסכמה מציגה מודל נתונים גרפי עבור נדל"ן, המבוסס על ישות "Property" (נכס) שמקושרת לישויות נוספות וליחסים ביניהן:**

**1. Property (נכס):**

**- מייצג את הנכס עם מזהה ייחודי `****transaction\_id`.**

**2. Duration (משך זמן):**

**- מייצג את משך הזמן באמצעות `tenure`.**

**- יש קשר `HAS\_DURATION` שמקשר את הנכס למשך הזמן, עם `start\_date` ו-`end\_date`.**

**3. Date of Transfer (תאריך העברה):**

**- מייצג את תאריך העברת הנכס באמצעות `date`.**

**- יש קשר `SOLD\_ON` שמקשר את הנכס לתאריך העברה, עם `transaction\_id` ו-`sale\_type`.**

**4. Old/New (ישן/חדש):**

**- מייצג את הסטטוס של הנכס באמצעות `status`.**

**- יש קשר `HAS\_STATUS` שמקשר את הנכס לסטטוס, עם `registration\_date`.**

**5. County (מחוז):**

**- מייצג את המחוז בו הנכס נמצא באמצעות `name`.**

**- יש קשר `PART\_OF` שמקשר את הנכס למחוז, עם `start\_date` ו-`end\_date`.**

**6. PPD Category Type (סוג קטגוריה של PPD):**

**- מייצג את קטגוריית ה-PPD באמצעות `category`.**

**- יש קשר `HAS\_CATEGORY` שמקשר את הנכס לקטגוריה, עם `reason`.**

**7. \*\*Town/City (עיר/יישוב)\*\*:**

**- מייצג את העיר או היישוב בו נמצא הנכס באמצעות `name`.**

**- יש קשר `LOCATED\_IN` שמקשר את הנכס לעיר או יישוב, עם `date\_from` ו-`date\_to`.**

**8. Property Type (סוג נכס):**

**- מייצג את סוג הנכס באמצעות `type`.**

**- יש קשר `HAS\_TYPE` שמקשר את הנכס לסוג הנכס, עם `sub\_type`.**

**9. Price (מחיר):**

**- מייצג את המחיר של הנכס באמצעות `amount`.**

**- יש קשר `HAS\_PRICE` שמקשר את הנכס למחיר, עם `transaction\_id`.**

**המודל מאפשר ניתוח נתונים מורכב על נכסים ואת הקשרים ביניהם, מספק תמונה מפורטת על תכני הנכסים וההיסטוריה שלהם במגוון קטגוריות.**