**README**

**שמות המגישים:** רתם קשאני 209073352 ודויד קופלב 208870279  
  
**רשימת הקבצים שהוגשו:**

1. **duckdb\_and\_sqlite.py**: בקובץ הזה אנחנו מגדירים את השם של הקובץ CSV שלנו, לאחר מכן אנחנו מתחברים לשרת של DUCKDB וטוענים ויוצרים טבלה מתוך המידע שנמצא בקובץ שלנו. לאחר מכן יש שאילתות על המידע שנמצא בתוך הטבלה ויוצרים DataFrame של pandas ובסיום היצירה של הDataFrames אנחנו סוגרים את החיבור לDUCKDB ופותחים חיבור לSQLite. לאחר מכן, אנו כותבים את הDataFrames לSQLite וסוגרים את החיבור.

2. **price\_paid\_records.db**: בקובץ זה נמצא הדאטה בייס הקטן (הDataFrames שכתבנו לSQLite בקובץ duckdb\_and\_sqlite.py). בנוסף קובץ זה מכיל בסיס נתונים SQLite עם נתוני נדל"ן מזויפים שנוצרו לצורך הדגמה והרצת הפרויקט. הוא כולל טבלאות עם מידע על נכסים, עסקאות, בעלים, וניתוחים סטטיסטיים שונים. בסיס נתונים זה משמש כתחליף לנתונים אמיתיים כאשר אין גישה למאגר הנתונים המקורי, ומאפשר להדגים את הפונקציונליות של המערכת ללא תלות בנתונים חיצוניים.

3. **price\_paid\_dashboard.py**: בקובץ זה אנחנו יוצרים אפליקציית Streamlit לתצוגת לוח מחוונים של נתוני נדל"ן. בהתחלה, אנחנו מתחברים לבסיס נתונים SQLite שמכיל את הנתונים ויוצרים תפריט צד לניווט בין עמודים שונים בלוח המחוונים. כל עמוד מציג תצוגה שונה של הנתונים - רשימת שאלות, סיפור ותובנות, שורות דוגמא מטבלאות בבסיס הנתונים, וגרפים המציגים עסקאות לפי חודש, מחיר ממוצע לפי סוג נכס ומצב, מחיר ממוצע לפי שנה ותקופות עם שינויים חדים במחירים. בסיום הפעלת האפליקציה, אנחנו סוגרים את החיבור לבסיס הנתונים.

4. **requirements.txt**: בקובץ זה נמצאים כל הספריות והחבילות שאנחנו צריכים להתקין כדי להריץ את הפרויקט שלנו. זה מכיל את הספריות duckdb, pandas, streamlit, matplotlib ו-numpy. הספרייה sqlite3 היא חלק מספריית הליבה של פייתון, ולכן אין צורך להתקין אותה באופן ספציפי.  
בנוסף, הוא מכיל את faker עבור חלק ג של התרגיל.

5. **תרגיל מתגלגל מסכם. docx**: קובץ זה מתאר בפירוט רב את התרגיל המסכם שביצענו. המסמך מספק תיאור מקיף של מאגר הנתונים שנבחר - מחירי נדל"ן בבריטניה, כולל מקורות המידע, התכנים שנכללו והמאפיינים הטכניים. הוא מתאר את המערכת התיאורטית שתוכננה לאיסוף, עיכול ועיבוד הנתונים. המסמך מציג את השאלות המרכזיות שנשאלו, התובנות שניתן לקבל מהן, והשאילתות שנכתבו כדי לענות עליהן. כמו כן, הוא מספר את הסיפור שרצו לספר דרך הניתוח - היסטוריית שוק הנדל"ן באנגליה וויילס בין 1995 ל-2017. המסמך מסביר את השיטות שנבחרו להקטנת הנתונים, את הוויזואליזציות הנבחרות והסיבות לבחירתן. הוא גם כולל הסברים איך להריץ את הקוד ולהציג את הדאשבורד, וצילומי מסך של הדפים העיקריים בדאשבורד. לבסוף, המסמך מציג ציור סכמתי של המערכת ומתאר את האחריות והעבודה של כל סטודנט במהלך התרגיל.  
בנוסף בקובץ זה יש את חלק ג, בחלק ג יש רשימה של nodes ושל relations שהגדרנו על מנת למדל את חלק מהעמודות ובסוף יש ציור שהוספנו שמציג את כל ה nodesוה relationsשהגשנו.

6. **spark\_property\_analysis.py**: בקובץ זה אנחנו מבצעים ניתוח על נתוני נדל"ן מזויפים באמצעות PySpark בסביבת Google Colab. הקוד מתחיל בהתקנת הספריות הנדרשות ויצירת סשן Spark. לאחר מכן, אנחנו טוענים קובץ JSON עם נתוני נדל"ן מזויפים ומבצעים עיבוד מקדים של הנתונים, כולל המרת טיפוסי נתונים. הניתוח העיקרי מתמקד בזיהוי 10 המחוזות עם מחירי הנדל"ן הממוצעים הגבוהים ביותר לכל שנה. התוצאות נשמרות בבסיס נתונים SQLite. בסיום, הקוד מאפשר הורדה של קובץ בסיס הנתונים וסוגר את סשן Spark. קובץ זה מדגים שימוש ב-PySpark לניתוח נתונים בקנה מידה גדול, כולל שימוש בפעולות מתקדמות כמו groupBy, window functions, ו-UDF (User Defined Functions).  
  
7. **fake\_propery\_data.py**: בקובץ זה אנחנו יוצרים נתוני נדל"ן מזויפים באמצעות ספריית Faker. הקוד מייצר 185,000 רשומות של נתוני נדל"ן, כל אחת מכילה מידע כמו מזהה עסקה, מחיר, תאריך העברה, סוג הנכס, סטטוס חדש/ישן, משך החזקה, עיר, מחוז, מדינה, סוג קטגוריה ומצב הרשומה. המחירים נקבעים באופן אקראי בין 50,000 ל-5,000,000 דולר. הנתונים המזויפים נשמרים בקובץ JSON בשם 'fake\_property\_data.json'. קובץ זה מאפשר לנו ליצור מסד נתונים גדול ומגוון לצורך בדיקות ופיתוח של המערכת שלנו.  
  
8. **fake\_property\_data.json**: קובץ זה מכיל את נתוני הנדל"ן המזויפים שנוצרו על ידי הסקריפט fake\_property\_data.py. הוא מאחסן את הנתונים בפורמט JSON, שהוא פורמט קל משקל לאחסון וחילוף נתונים. כל רשומה בקובץ מייצגת עסקת נדל"ן בודדת ומכילה שדות כגון מזהה עסקה, מחיר, תאריך, סוג נכס, מיקום וכו'. הקובץ משמש כמקור נתונים לניתוח ב-PySpark ולהדגמת פונקציונליות המערכת. השימוש בנתונים מזויפים מאפשר לנו לעבוד עם כמות גדולה של נתונים מבלי להשתמש בנתונים אמיתיים ורגישים. הקובץ שסיפקנו מכיל רק 50 רשומות ראשונות, בעוד שהקובץ המקורי היה בגודל של 71MB והכיל 185,000 רשומות.

9.**Price Paid Records Table Summary.docx**: קובץ זה מכיל סיכום של הטבלאות בבסיס הנתונים SQLite שלנו. הוא מתאר שש טבלאות עיקריות:

1. עסקאות לפי חודש: מציגה את התפלגות העסקאות לאורך חודשי השנה.

2. מחיר ממוצע לפי סוג ומצב: מראה את הבדלי המחירים בין סוגי נכסים שונים ובין נכסים חדשים לישנים.

3. מחיר ממוצע לפי שנה: עוקבת אחר מגמות מחירי הנדל"ן משנת 1995 עד 2017.

4. שינויי מחירים: מדגישה תקופות עם שינויים משמעותיים במחירי הנדל"ן.

5. הערים המובילות בעליית מחירים (1995-2017): מזהה את הערים עם הצמיחה המשמעותית ביותר במחירי הנדל"ן.

6. מחיר ממוצע לפי סוג נכס ומחוז: מספקת ניתוח מפורט של מחירי נדל"ן באזורים שונים ולפי סוגי נכסים.

קובץ זה מסייע בהבנת מבנה הנתונים ומטרת כל טבלה בניתוח שלנו.

10. **price\_paid\_records.csv**: הקובץ מכיל מידע על עסקאות רכישת נכסים באנגליה וויילס, עם השדות הבאים: מזהה ייחודי לעסקה, מחיר המכירה, תאריך המכירה, סוג הנכס, גיל הנכס, סוג הבעלות, שם העיר או היישוב, מחוז או מחוזה, מדינה, סוג העסקה, ומצב הרשומה בקובץ החודשי (הוספה, שינוי, מחיקה). הקובץ שסיפקנו מכיל רק 50 שורות ראשונות, בעוד שהקובץ המקורי היה בגודל של 2.24GB והכיל 22489348 שורות.