**first de project**

**עמית גרומט**

**דאטה**

מידע על השכרות דירות במספר ערים באירופה דרך הפלטפורמה Airbnb המידע מוחלק למספר טבלאות, לכל עיר יש שתי טבלאות שונות אחת לסופי שבוע ואחת לאמצע שבוע. סה"כ יש 16 טבלאות בכל אחת אלפים בודדים של רשומות ו12 פיצ'רים רלוונטיים. [קישור לדאטה באתר Kaggle.](https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities) המחיר הוא לשתי אנשים לשתי לילות ביורו.

פיצ'רים בוליאניים: נכס פרטי או משותף, נכס ברשות עסק, האם הנכס של משתמש פרמיום.

פיצ'רים מספריים: מחיר, מספר חדרים, דירוג ניקיון , דירוג כללי, מרחק ממטרו, מרחק ממרכז העיר, גודל דירה, רמת התעניינות בעיר, רמת המסעדות בעיר.

**טבלאות מחושבת עצמאית**

בנוסף למידע הקיים בטבלאות ברצוני להוסיף עוד מספר טבלאות שאחשב עצמאית.

1. טבלאות weekEnds שמכילה מידע על כלל הדירות בסופי שבוע.
2. טבלאות weekDays שמכילה מידע על כלל הדירות באמצע שבוע.
3. מספר טבלאות מסוג pivot tables שעוד לא החלטתי בדיוק מאיזה סוג הם יהיו (נדרש עוד מחקר על הדאטה בשביל להבין איזה סוג מידע כדאי לחשב ולהציג).
4. טבלת חישוב פונקציית המחיר: טבלה מרכזית שמחזיקה קישור לכל שאר הטבלאות מכילה את העמודות: שם הטבלה, מספר רשומות בטבלה, ואת פונקציית החישוב הלינארית לחישוב מחיר בהתבסס על הפיצ'רים שתחושב בזמן אמת על ידי מודל רגרסיה לינארית אחת לשלושה ימים.
5. לאחר חישוב המחירים נבצע קלאסטרינג בעזרת knn וניצור טבלאות לכל קלאסטר.
6. יצירת טבלת של דירות בעזרת ערך unique שמיוצר בעזרת dist עד דיוק של 5 נקודות אחרי הנקודה.

**אופטימצזיות**

כדי לדמות מערכת גדולה עם המון מידע אבצע מספר אופטימזציות כדי לנסות ליעל את השמירה מבחינת אחסון.

1. במקום לשמור מידע יחסי כמו מרחק ממטרו גודל דירה וכו כמספר נבצע binning ונרמול (בנוסף יכול ליעל חישוב רגרסיה לינארית וקלסטרינג).
2. נחפש אנומליות לדוגמא מחיר או גודל דירה גבוה משמעותית מהממוצע ונמחק אותם מהדאטה סט.
3. מחיקה של מידע לא רלוונטי כגון גובה ורוחב (קואורדינטות), האם לבעל הדירה יש 2-4 נכסים או 4+ נכסים והכנסה תחת קטגוריה אחת של עסק.
4. one-hot-encoding לפיצ'רים room\_type room\_shared room\_private ושמירה של הערכים המקוריים ומיפוי שלהם בטבלה לצורך decoding.
5. הורדת רשומות עם ערכי null או ערכים לא unique בשדה dist. מכיוון שיש מספיק דאטה, שיטות כמו ממוצע של כלל הערכים וכו' יכולים לחוטא למטרה מכיוון שכל פרמטר פה חשוב וממוצע של ערכים בשדה מסוים אינו מעיד על הערך שכנראה היה שם לדוגמא דירה זהה לגמרי אבל בעלת 10 חדרים ולא 2 תהיה בעלת מחיר שונה לגמרי.

**חישוב פונקציית מחיר**

נשתמש במודל רגרסיה לינארית למטרת חישוב המחיר נכוון לmse של 900 שבעצם אומר שאנו טועים בממוצע ב30 יורו מתורגם ל6 אחוז שגיאה ממוצעת.

A graph showing a couple of circles

Description automatically generated with medium confidence

A close-up of a graph

Description automatically generatedA close-up of a graph

Description automatically generated

A pie chart with numbers and a number of percentages

Description automatically generated with medium confidence

**סיכום**

המערכת מכילה מספר מיקרו סרוויסים ועובדות עם קאפקא וpostgres.

ישנם מספר דברים שצריך להוסיף כמו data Lake יותר גדול ומורכב ומעבר לדוקר וs3.

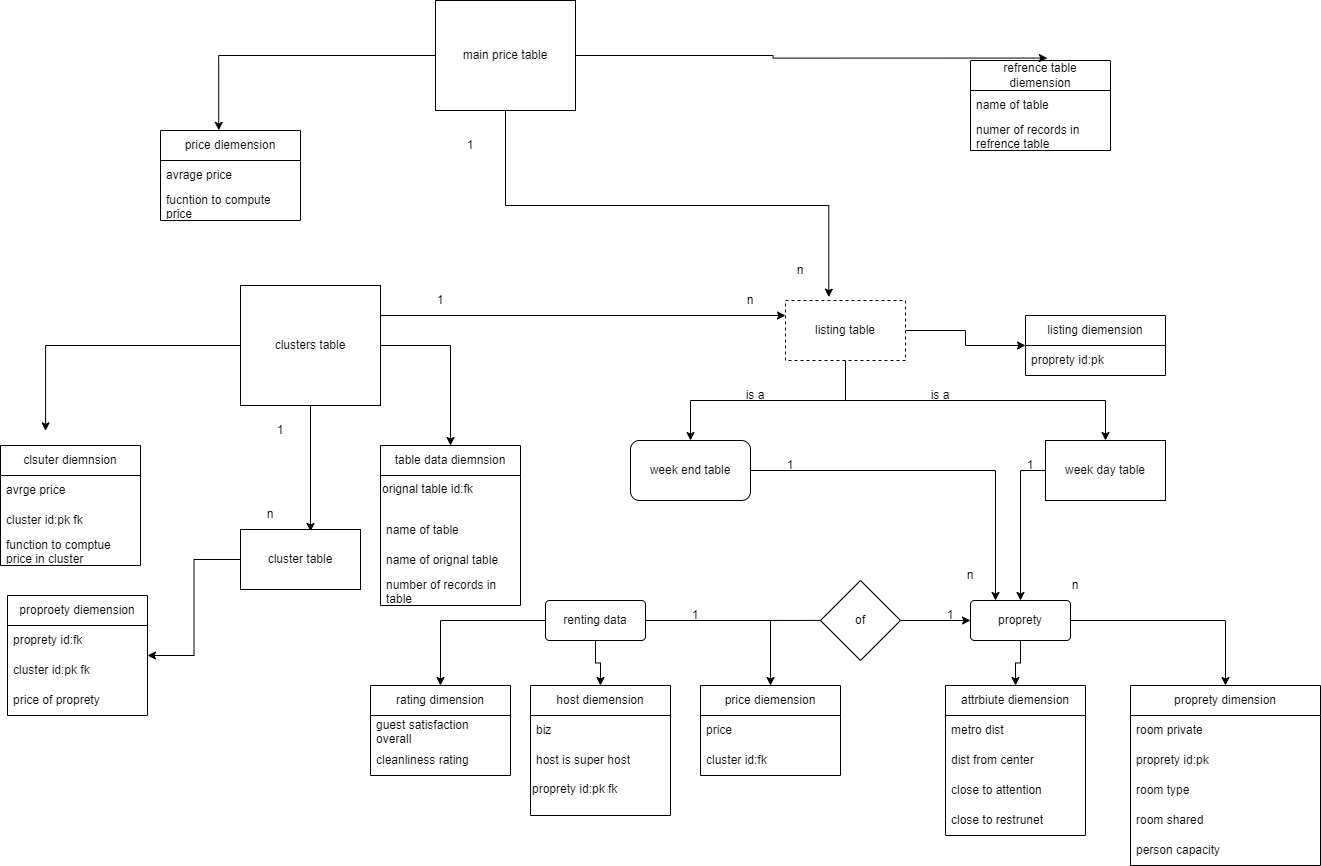
תוצאות החיזוי של המודל לא מוצלחות במיוחד אנו עומדים על בערך 17 אחוז טעות(mse חלקי ממוצע מחיר בריבוע). במידה ומנקים מהמידע דירות עם מחיר גבוה מעל פי שתיים מהממוצע הטעות יורדת לרמות של 7 אחוז. ההשערה שלי היא שדירות שמחיריהם יקר פי שתיים מהממוצע מציעות משהו שלא מוצג בדאטא לדוגמא נוף יפה או עיצוב פנים מרשים. בנוסף ישנם ערים בהם המודל ממש לא הצליח לדוגמא פריז או אמסטרדם השערתי היא שמידע שם לא מדויק או שדורש עבודת מחקר גדולה יותר שזה כמובן לא מטרת המשימה.

למדתי:

* עבודה עם orm מסוג sqlalchemy.
* שימוש רחב מאוד בkafka והבנה של המערכת.
* עבודה בתצורה של מיקרו סרווסים לעומת oop.

הערות:

המערכת לא מוכנה לרמת "פרודקשן" לאחר ההערות והpr יבצעו שינויים כדאי לתמוך יותר בקלין קוד המערכת הזאתי היא כרגע ברמת v1.

****