מחסני נתונים - פרויקט Airbnb

מבוא

חברת Airbnb בעלת אתר אינטרנט המתווך בין שוכרי יחידות נופש להשכרה או למגורים לטווח קצר לבין בעלים של יחידות כאלו. הבעלים (hosts) ברשותם לפחות יחידת נופש או דיור אחת (listings) שמיועדת לאירוח אנשים שונים. האורחים יכולים להזמין את היחידה ולאחר סיום שהותם באפשרותם לכתוב ביקורת. נתונים לגבי פעילויות באתר Airbnb נאספו בתאריך 7/12/2018 אודות זמינות ומשובים עבור יחידות נופש ומגורים[1].

לרשותכם (ב-Moodle) קבצי נתונים מעובדים שמופיעים עם מידע אודות יחידות נופש ודיור במלבורן, אוסטרליה. עליכם לבנות מחסן נתונים ומערכת BI סביב תהליך גרעיני שתבחרו.

אוסף הנתונים ניתן בפורמט CSV במספר קבצים:

- listings – מידע על יחידות נופש ודיור
- calendar – מידע על זמינות ומחיר לכל יחידת נופש ודיור מתאריך 8/12/2018 עד 7/12/2019
- reviews – מידע על המשובים שנתנו ליחידות הדיור על ידי משתמשים

תיעוד מלא של מבנה הקבצים ניתן למצוא בנספחים.

[1] מקור הנתונים: https://www.kaggle.com/tylerx/melbourne-airbnb-open-data/activity

דרישות הפרויקט

עיצוב ובניית מחסן נתונים וביצוע תהליך ETL מקבצי הנתונים הקיימים, לאחר מכן להשתמש במחסן הנתונים לצורך הפקת דו״ח אנליטי.

חלק א: תכנון ו-ETL

סעיף 1: עיצוב מחסן נתונים בסכמת **כוכב** על פי השלבים שנלמדו בכיתה:

א. זיהוי התהליך (Business Process) בו מחסן הנתונים מתמקד
ב. בחירת הגרעין (Grain)
ג. בחירת מימדי מחסן הנתונים
ד. זיהוי העובדות

יש לבחור תהליך עסקי כללי במערכת, שרלוונטי עבור מספר רב של עסקים או משתמשים. יודגש כי עיצוב מחסן הנתונים מבוסס על התוצרים הנדרשים ממנו. לכן, יש לתכננו כך שתוכלו להפיק ביעילות את הדו״חות לניתוח שתכננתם. כמו כן – שימו לב כי לא כל הנתונים המקוריים רלוונטיים לתהליך שתבחרו.

תפוקות נדרשות:

פירוט (בקובץ PDF) של העיצוב מחסן הנתונים, תרשים קונספטואלי של מחסן הנתונים, כולל הגדרת השדות בפורמט המתואר בדוגמה:

דוגמה - טבלת מימד יחידות השכרה (Listings)

שדה	טיפוס נתונים	מקור נתונים	הערות
listing_id	int	שדה id בקובץ listings	מפתח ראשי, ייחודי
avg_cost	double	מחושב ע״י AVG(price) מטבלת calendar.	מהווה תמונת מצב קיים, יש לעדכן בהינתן נתונים חדשים.

סעיף 2: ביצוע תהליך ETL למידע הקיים:

בהתאם לעיצוב בסעיף 1 עליכם לבנות תהליך ETL בעזרת Python המייצר מסלול קריאת הנתונים (Pipeline) מקבצי הקלט (Extract), עיבוד הנתונים (Transform) וטעינתם לבסיס הנתונים MySQL (Load).

דגשים לשלב זה:

א. עליכם לבנות תהליך אוטומטי ככל הניתן, כך שיהיה ניתן לשחזר אותו (או להוסיף נתונים חדשים) בקלות, לכן הימנעו ככל הניתן מעריכת הנתונים ידנית באקסל או שינויים נקודתיים שאינם חלק מהקוד שתגישו.
ב. יש לתעד היטב את קוד ה-Python ו-SQL לעיבוד הנתונים וטעינתם לבסיס הנתונים.

ג. נדרש להשתמש בחבילת Pandas לצורך ביצוע תהליך ETL.

ד. הוסיפו מספר שלבי בדיקת נתונים, הן לצורך הימנעות מרשומות זבל וכפילות ברשומות, והן לבדיקת תקינות התהליך בסופו.

תפוקות נדרשות:

- פירוט קוד Python ושאילתות SQL לביצוע התהליך (בקובץ Jupyter notebook), כולל תיעוד הפעולות והשיקולים שנלקחו במהלך העבודה.
- מידע על הנתונים בכל טבלה במחסן הנתונים – כמות שדות, כמות רשומות, טווחי ערכים רלוונטיים(למשל טווח תאריכים). יש לצרף דוגמית של נתונים (כ- 5-10 שורות) מכל טבלה.

חלק ב׳: יצירת דוחות

דו״ח אנליטי המתבסס על מחסן הנתונים שבניתם. בחלק זה עליכם להציג שאלת מחקר, לנסח שאילתות לתשאול הנתונים, לבצע ניתוח ולהציג את המסקנות.

תפוקות נדרשות: קובץ PDF המציג את שאלת המחקר, תהליך הניתוח (יש לצרף קוד Python) ואת המסקנות העולות ממנו.

נהלים והנחיות כלליות

1. ציון יחושב באופן הבא:
   - זיהוי תהליך עסקי ועיצוב מחסן הנתונים
   - תהליך ה-ETL
   - נכונות שלבי טעינת המידע
   - רמת האוטומציה בתהליך
   - רמת התיעוד של תהליך ETL
   - בחירת שאלת מחקר, מקוריות ומורכבות העבודה
   - איכות הקוד הטעינה וניתוח הנתונים
   - הצגת מסקנות (מילולית וויזואלית)
2. הגשה בקבוצות של 3 סטודנטים.
3. מועד הגשה מפורסם באתר הקורס.
4. ההשגה תיעשה על ידי סטודנט אחד מחברי הקבוצה בקובץ ZIP הכולל את:
   א. קובץ PDF המכיל את תוצרי חלק א׳ (סעיף 1) ו-ב׳ (ללא קוד).
   ב. קובץ Jupyter Notebook המכיל את הקוד ששימש ליצירת מחסן הנתונים (חלק א׳ סעיף 2).
   ג. קובץ Jupyter Notebook המכיל את הקוד ששימש להכנת הדוחות והגרפים ממחסן הנתונים (חלק ב׳ – קוד Python).
5. כל הקבצים צריכים להכיל את ת.ז של חברי הקבוצה ומס׳ הקבוצה. נא לא לצרף את קבצי הנתונים.
6. חריגה מפורמט ההגשה (בפרט תיעוד לקוי, קוד שלא ניתן להרצה או מחסור בקבצי התקנה נלווים) כמו גם איחור במועד ההגשה יובילו להורדת ציון.
7. המסמך יהיה כתוב בגופן David, גודל 12, עם מרווח של שורה וחצי.

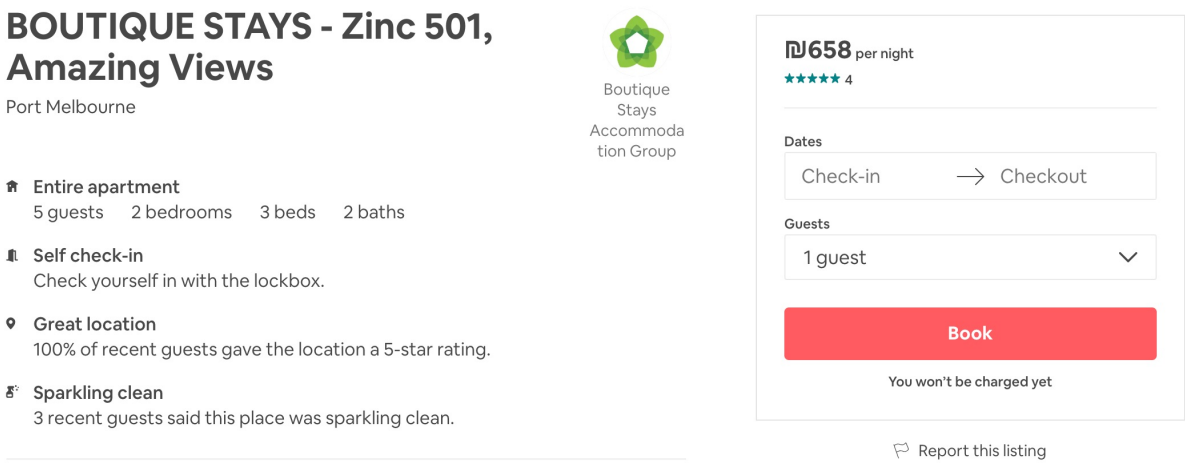
בהצלחה !

נספחים

נספח א׳: קובץ listings.csv

מרבית השדות מתייחסות לנתונים שנאספו מדף האינטרנט של יחידת הדיור (listings).

לדוגמה: עבור id=464528 עמוד יחידת הדיור https://www.airbnb.com/rooms/464528 נראה כך:



כאשר ניגש לנתונים אפשר לראות ערכים שמופיעים באתר:

```
In [65]: listings[listings.id == 464528][['id','name','guests_included','bathrooms','beds','price','amenities','description']]
Out[65]:
```

	id	name	guests_included	bathrooms	beds	price	amenities	description
138	464528	BOUTIQUE STAYS - Zinc 501, Amazing Views	4	2.0	3.0	$325.00	{TV,"Cable TV",Internet,Wifi,"Air conditioning...	Professionally managed by Boutique Stays, Zinc...

שימו לב שהמרה מדולר אוסטרלי לשקל. כמובן יכולים להיות שינויים קלים מכיוון שהנתונים נאספו ב-7/12/2018.

רשימת כל השדות הקיימים בקובץ listings:

Field	Description
id	Listing ID on Airbnb site.
listing_url	Listing url from Airbnb.
scrape_id	Ignored[2]
last_scraped	Ignored
name	The name of the listing. [3]
summary	The Summary section of the listing ad. A paragraph highlighting the listing.
space	The Space section of the listing ad. A paragraph describing the listing space.
description	The Description section of the listing ad. A paragraph describing the listing details.
experiences_offered	Any touring services, or Experiences, offered by the host.

[2] אתם מתבקשים לא להשתמש בשדות שבתיאור שלהם מופיע Ignored. שאר השדות השימוש בהם לשיקולכם.
[3] תיאור שצבעו כחול הכוונה שעמודות מוגדרות כטקסט חופשי ובמסגרת הפרויקט לא נדרש לבצע ניתוח טקסט כלשהו, למעט האם קיים טקסט או לא.

neighborhood_overview	The Neighborhood section of the listing ad. A paragraph describing the listing neighborhood.
notes	The Notes section of the listing ad. A paragraph listing the info that the host wanted the guests to know.
transit	Transportation info on the listing ad.
access	A paragraph describing what areas of the listing the guests can acces.
interaction	A paragraph describing how the host will interact with guests.
house_rules	A paragraph listing the house rules.
thumbnail_url	Ignored
medium_url	Ignored
picture_url	Ignored
xl_picture_url	Ignored
host_id	Host ID on Airbnb site.
host_url	Url of the host profile page.
host_name	Host name on the listing.
host_since	The date when the host started.
host_location	Host location from the profile.
host_about	A paragraph written by the hosts about themselves.
host_response_time	Average time the host responses to a guest inquiry: "within an hour", "within a few hours", "within a day", "within a few days', etc.
host_response_rate	The percentage that the host response to all guest inquiries.
host_acceptance_rate	The percentage that the host accept the guests request to stay.
host_is_superhost	Whether the host is a superhost: t or f.
host_thumbnail_url	Url of the thumbnail host profile photo.
host_picture_url	Url of the host profile photo.
host_neighbourhood	The neighborhood host lives.
host_listings_count	How many listings the host has.
host_total_listings_count	How many listings the host has.
host_verifications	Verifications include email, phone, government ID, etc.
host_has_profile_pic	Whether the host has uploaded a photo to the profile: t or f.
host_identity_verified	Whether the host's ID is verified: t or f.
street	Ignored
neighbourhood	Ignored
neighbourhood_cleansed	Ignored
neighbourhood_group_cleansed	Ignored
city	Ignored
state	Ignored
zipcode	Ignored
market	Ignored
smart_location	Ignored

country_code	Ignored
country	Ignored
latitude	Latitude of the listing.
longitude	Longitude of the listing.
is_location_exact	Whether the listing location showing on the map of the ad is exact or not: t or f. Airbnb allows hosts to choose if they want to show the location exactly on the map or just approximately on the map.
property_type	The property type of the listing: apartment, house, etc.
room_type	Room type: entire apartment, private room, shared room, etc.
accommodates	The max number of guests the listing accommodates.
bathrooms	How many bathrooms. 0.5 means just a toilet and sink only, or a bathtub or shower only.
bedrooms	How many bedrooms.
beds	How many beds.
bed_type	What the type of the bed: full bed, sofa bed, etc.
amenities	Ignored
square_feet	The size of the property.
price	How much for a night, excluding cleaning and other fees.
weekly_price	How much for a week, excluding cleaning and other fees.
monthly_price	How much for a month, excluding cleaning and other fees.
security_deposit	How much for the security deposit.
cleaning_fee	How much for the cleaning fee.
guests_included	How many guests can be accommodated without extra fees.
extra_people	How much per person for extra guests.
minimum_nights	Minimum nights guests can book, if any.
maximum_nights	Maximum nights guests can book, if any.
calendar_updated	When is the last calendar update/modify made by host.
has_availability	Whether is available. 100% true as the listing won't be showing on the site otherwise.
availability_30	How many days available in the next 30 days.
availability_60	How many days available in the next 60 days.
availability_90	How many days available in the next 90 days.
availability_365	How many days available in the next 365 days.
calendar_last_scraped	Doesn't have much value.
number_of_reviews	How many reviews by previous guests.
first_review	The date of the first review.
last_review	The date of the latest review.
review_scores_rating	
review_scores_accuracy	
review_scores_cleanliness	
review_scores_checkin	

review_scores_communication	
review_scores_location	
review_scores_value	
requires_license	
license	
jurisdiction_names	
instant_bookable	Whether can be booked instantly: t or f.
is_business_travel_ready	t or f.
cancellation_policy	What is the cancellation policy: strict, moderate, flexible, etc.
require_guest_profile_picture	Whether requires guest profile photo upon booking.
require_guest_phone_verification	Whether requires guest phone number upon booking.
calculated_host_listings_count	How many listings the host have based on all listings scraped.
reviews_per_month	How many reviews received per month.

הערות:

- 22,895 רשומות.
- השדה id ייחודי ומופיע כמספר הרשומות.
- השדה host_id מייצג את הבעלים של הנכס. לכל host_id יש לפחות נכס אחד (id).
- הקובץ מכיל מידע לגבי הנכס, הבעלים של הנכס ומידע סיכומי עבור זמינות יחידת הדיור, דירוגים ומשובים.
- ישנן 96 עמודות! לשיקולכם אילו שדות להשתמש על מנת לענות על שאלת המחקר שלכם.
- כאשר המילה **ignored** מופיעה בתיאור (description) אתם מתבקשים לא להשתמש בשדה.
- לא נדרש לעשות ניתוח טקסט לשדות שיש בהם טקסט חופשי (מסומן בכחול). אבל, במידת הצורך - אפשר לבדוק האם קיים טקסט או לא?

נספח ב׳: קובץ calendar

הקובץ מייצג מועדים פנויים לביצוע הזמנה מ-7/12/2018 עד 6/12/2109. כאשר כל listing_id חוזר על עצמו 365. אם בשדה available מסומן t אז יחידת הדיור פנויה, אחרת מסומן f. השדה price בטבלה מייצג את המחיר עבור יחידת הדיור הזמינה.

רשימת כל השדות הקיימים בקובץ calendar:

Field	**Description**
listing_id	Listing ID on Airbnb site.
date	All dates on the next year or so.
available	Whether the listing is available: t or f.
price	The price for that date, if available.

הערות:

- 8,356,675 שורות! נקודה למחשבה – האם כל השורות נדרשות? בעיקר תלוי בשאלת מחקר שלכם.
- קיים קשר בין עמודת available ו-price כאשר יחידת הדיור זמינה קיים המחיר, אם היחידה אינה זמינה לא קיים המחיר.

נספח ג׳: קובץ reviews.csv

הקובץ מכיל משובים שאורחים כתבו ליחידת דיור.

Field	Description
listing_id	Listing ID on Airbnb site.
id	Review ID.
date	The date of the review posted.
reviewer_id	Guest ID on Airbnb site.
reviewer_name	Guest's first name.
comments	Comment contents

הערות:

- 486,920 משובים (רשומות) עבור 17,653 יחידות דיור שונות ומ-391,061 אורחים שונים.
- התאריך הראשון למשוב הינו 4/8/2010 והאחרון 7/12/2018 (תאריך איסוף הנתונים).