

# Personalize Hotel Booking

Analyzing Hotel Reviews over Booking.com

Noa Barbiro  
Yanir Calisar  
Kobi Shamama

kaggle™

Booking.com

yelp\*

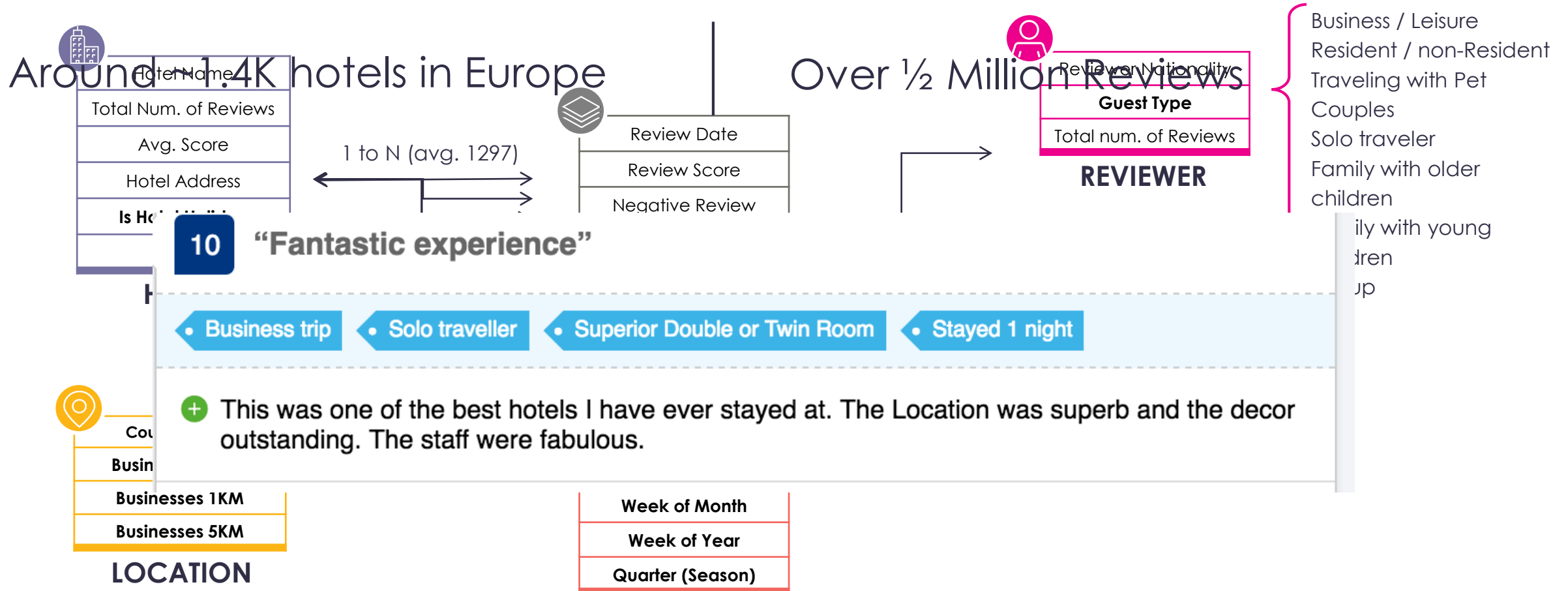
Google Maps





SO, WHAT DO WE HAVE HERE?

Booking.com



Published Dataset

1

## Enriched Hotel Reviews Dataset



Yanir • last updated 2 months ago

Overview

Data

Kernels

Discussion

Activity

Download (55 MB)

New Kernel

1 Files (54.55 MB)

hotel\_reviews\_en...

Download All

## hotel\_reviews\_enriched.csv

298 MB • Updated 2 months ago

Download

About this file

1. Enriched the original file with data from Yelp about nearby businesses (in Businesses\_100m, Businesses\_1km, Businesses\_5km)
2. Broken the Address field into Hotel\_Country, Hotel\_State, Hotel\_City
3. Broken the Tags field into Room\_Type, Room\_Type\_Level, Bed\_Type, Guest\_Type, Trip\_Type and Stay\_Duration
4. Broken the Review\_Date field into Day\_of\_Week, Day\_of\_Year, Days\_Since\_Review, Week\_of\_Month, Week\_of\_Year and Quarter\_of\_Year.
5. Used the jollyday DB to find if there is a holiday at the hotel country or the reviewer's country on the review date (Is\_Hotel\_Holiday, Is\_Reviewer\_Holiday)
6. Added Review\_Is\_Positive and Review\_Positivity\_Rate fields based on a comparison between the fields Review\_Total\_Negative\_Word\_Counts and Review\_Total\_Positive\_Word\_Counts

[Preview \(first 100 rows\)](#)[Column Metadata](#)[Column Metrics](#)

id	Hotel_Name	Hotel_Address	Hotel_Country	Hotel_State	Hotel_City	Hotel_lat	Hotel_lng
13	Hotel Arena	s	NL	NH	Amsterdam	52.3605759	4.9159683

kaggle™

Contributing Back to the Community

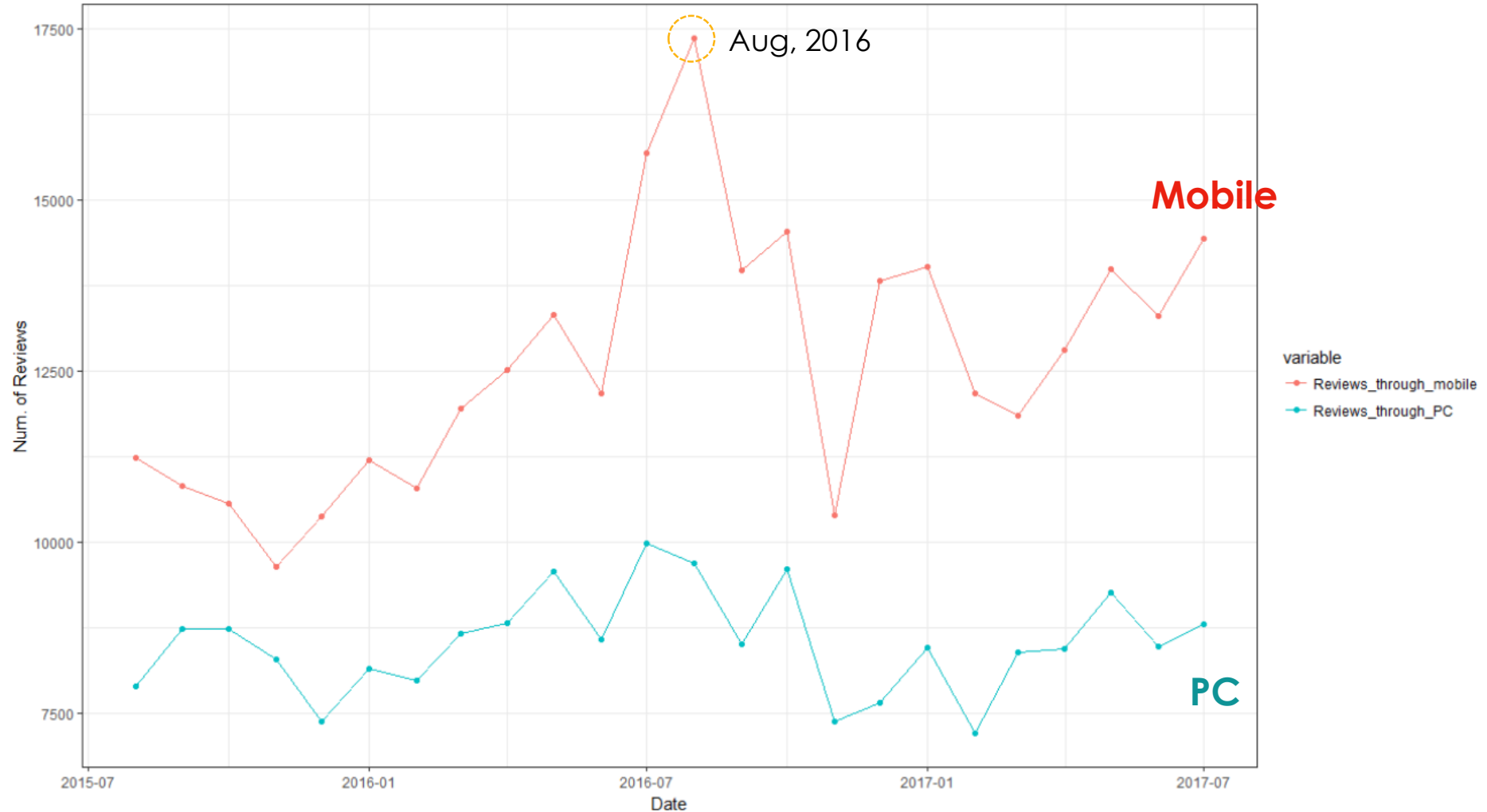


# SOME COOL STATS

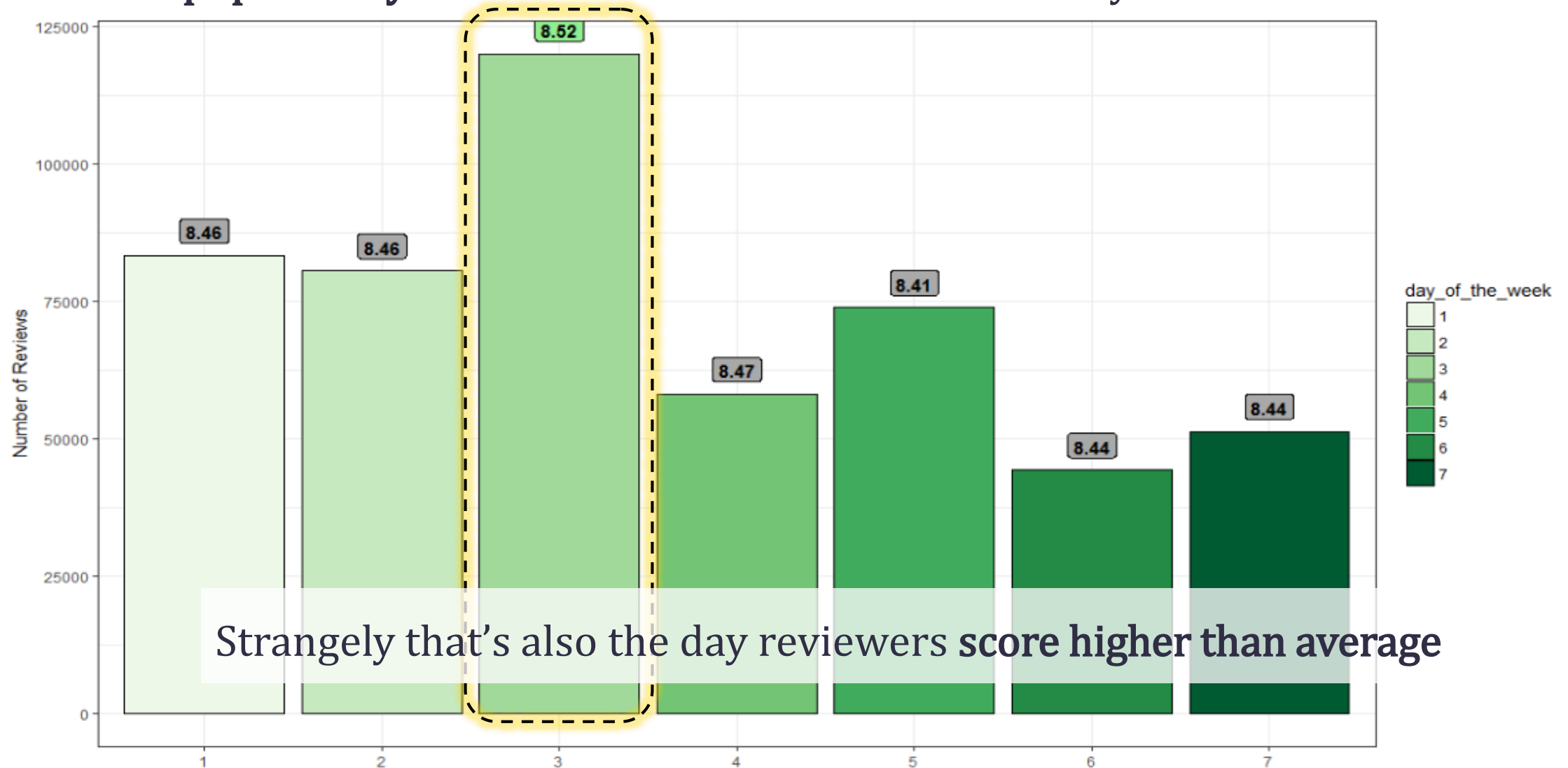


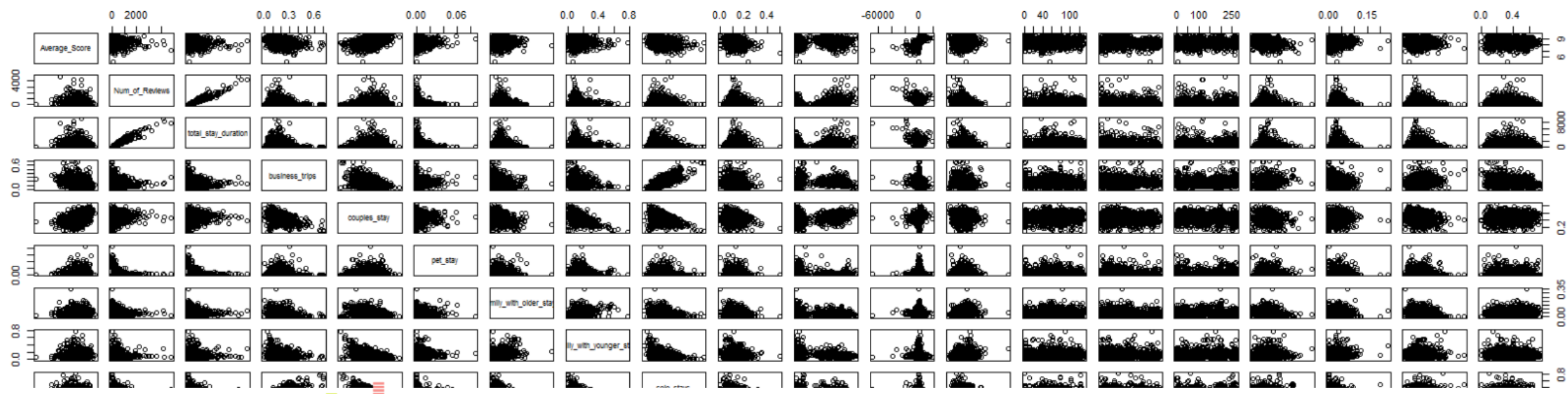
## EDA IN ACTION. FINDINGS

Over time **mobile** is used more than **pc** to submit reviews, almost ~200% more

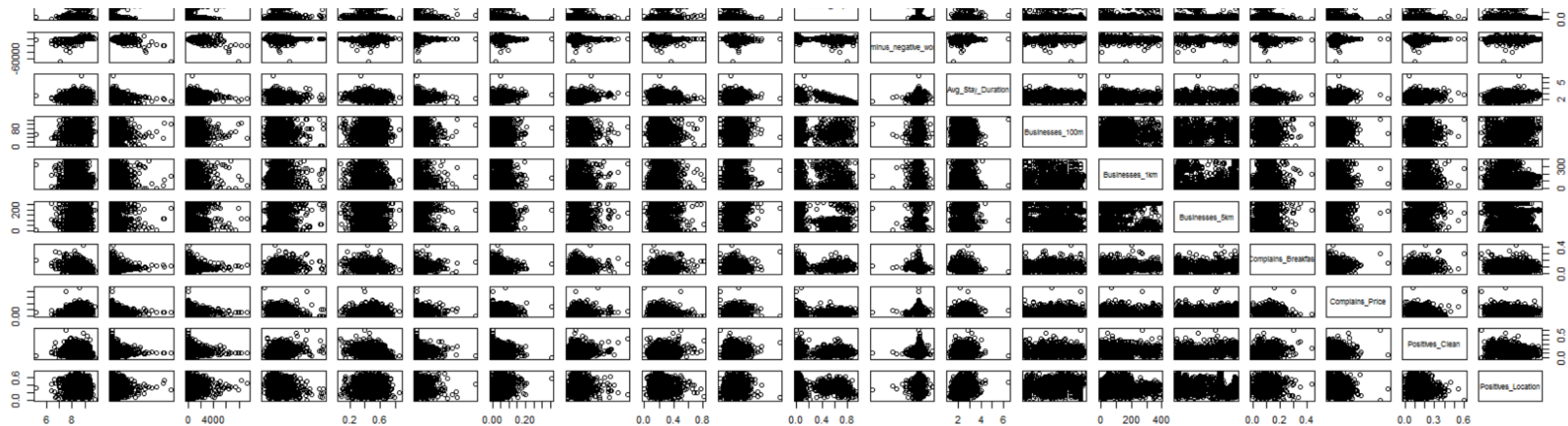


The most popular day for submission of a review is Tuesday

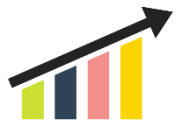




# INTERESTING CORRELATIONS



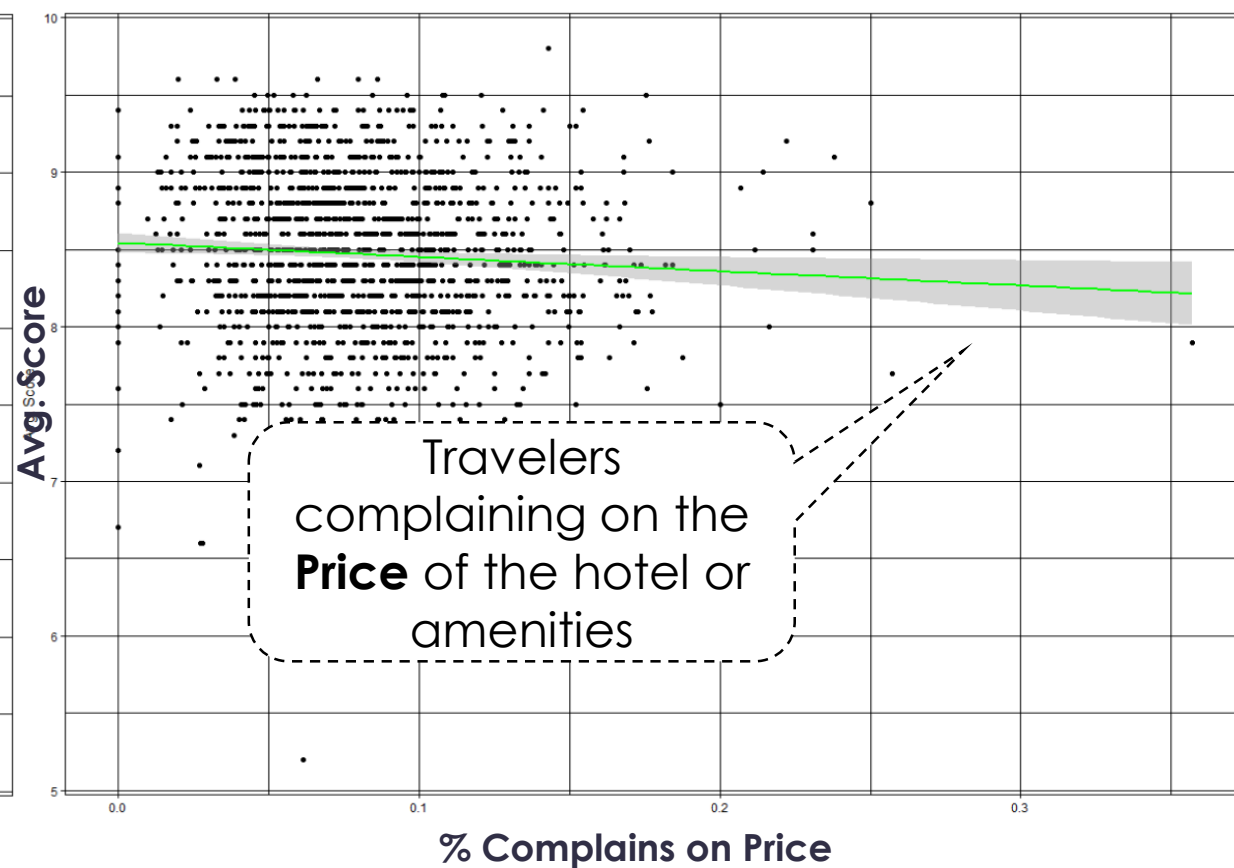
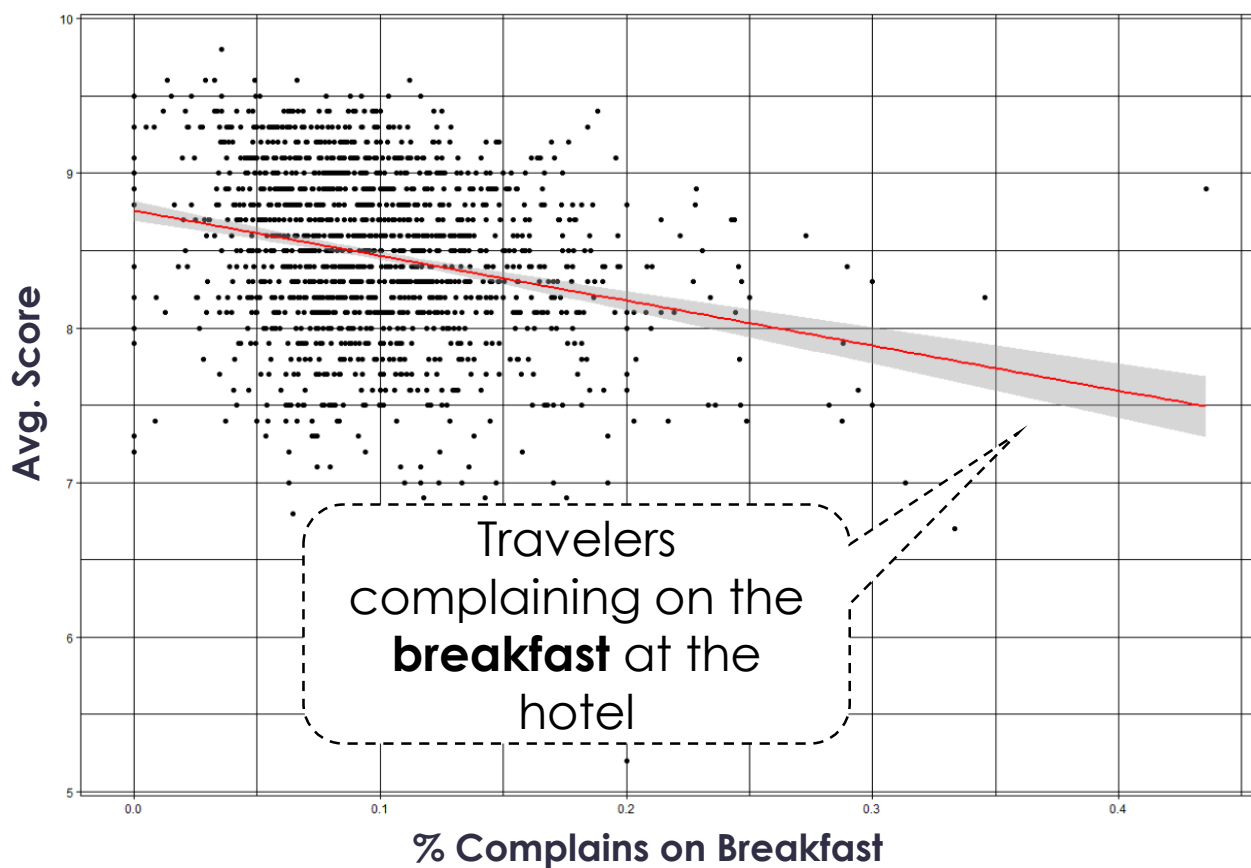


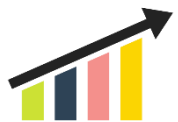


## CORRELATION & REGRESSION. FINDINGS

Bad Breakfast tends to impact review score

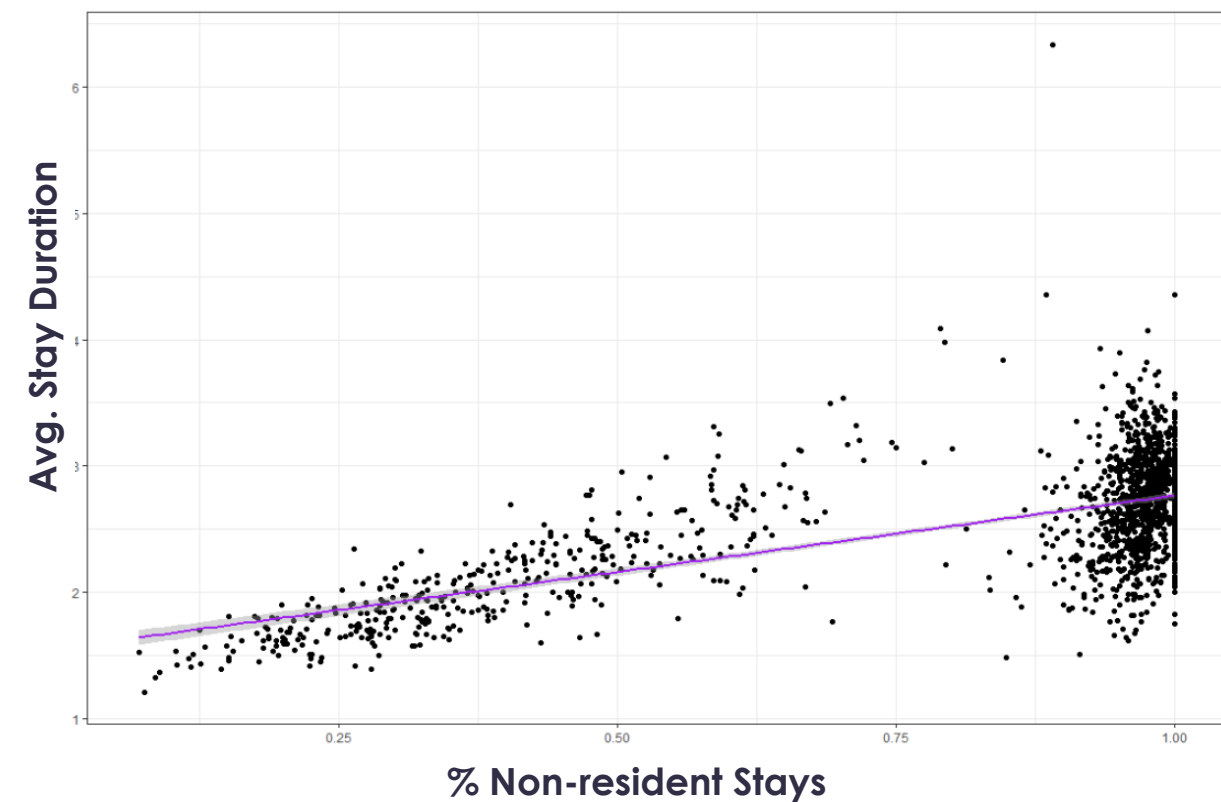
.. But not so much for **high price** !



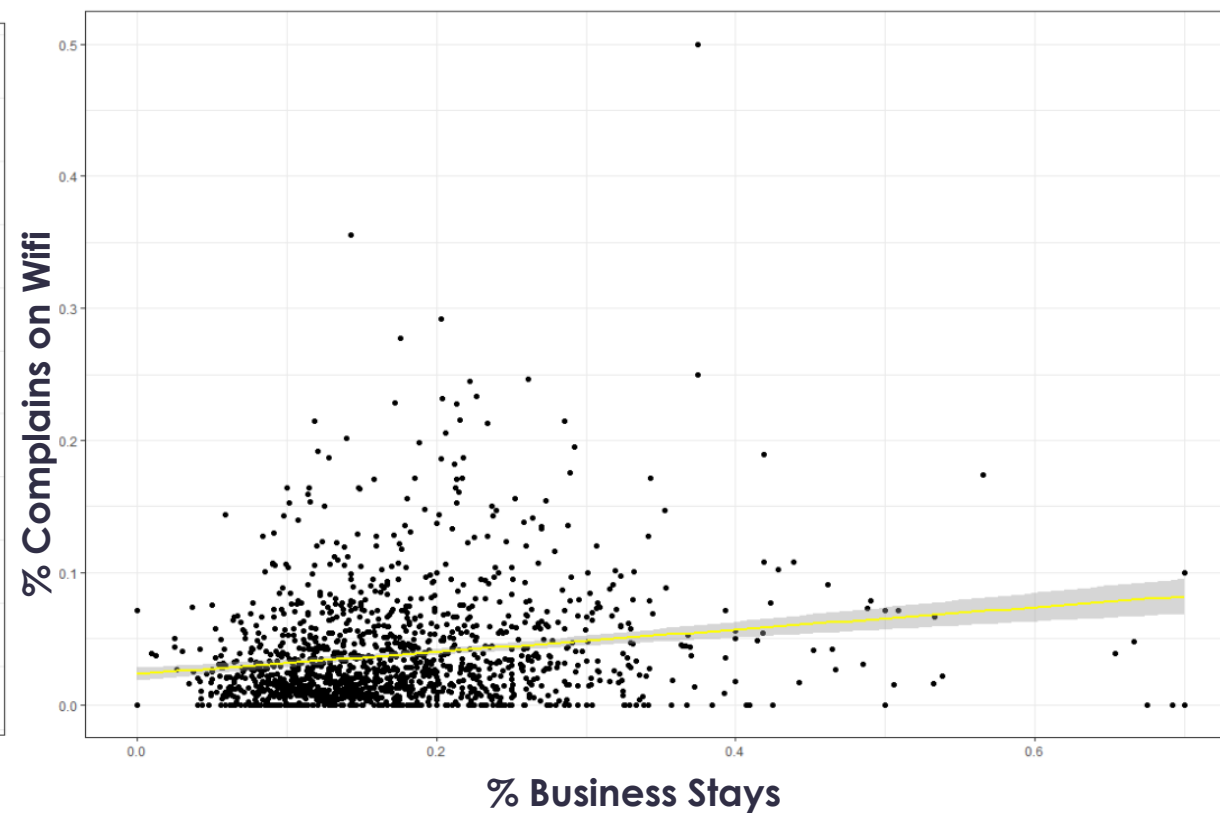


## CORRELATION & REGRESSION. FINDINGS

**Non-Resident Travelers correlate with longer avg. stay duration**



**Business Travelers correlate with higher rate of complains on Wifi**





# UNDERSTANDING THE LANDSCAPE **Booking.com**

**KAYAK**<sup>®</sup>

**Hotels.com**<sup>®</sup>

 **travelocity**<sup>®</sup>

 **Expedia**<sup>®</sup>

 **airbnb**

**agoda.com**  


**trivago**

 **tripadvisor**<sup>®</sup>

**lastminute.com**

# WHAT IS SO UNIQUE WITH AIRBNB ?

## Machine learning is driving growth at Airbnb



Stephanie Pandolph

Jun. 16, 2017, 10:43 AM 🔥 4,571



FACEBOOK



LINKEDIN



TWITTER



EMAIL



PRINT

This story was delivered to BI Intelligence "E-Commerce Briefing" subscribers. To learn more and subscribe, please [click here](#).

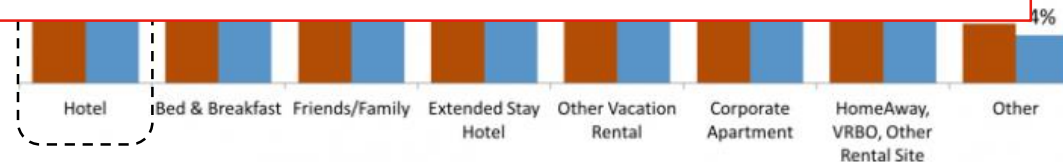
Machine learning (ML) has had a "profound" effect on Airbnb's business growth, according to the

### Where Global Airbnb Users Are Switching From



- The company uses a machine-learned search ranking model to personalize results for guests. The model factors in guests' tendencies

specific types of décor in places they book. The company feeds more than 100 characteristics into the model, which then uses the data to identify patterns and personalize search rankings.



Source: Morgan Stanley

BI INTELLIGENCE



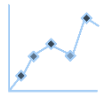


BUSINESS OBJECTIVES

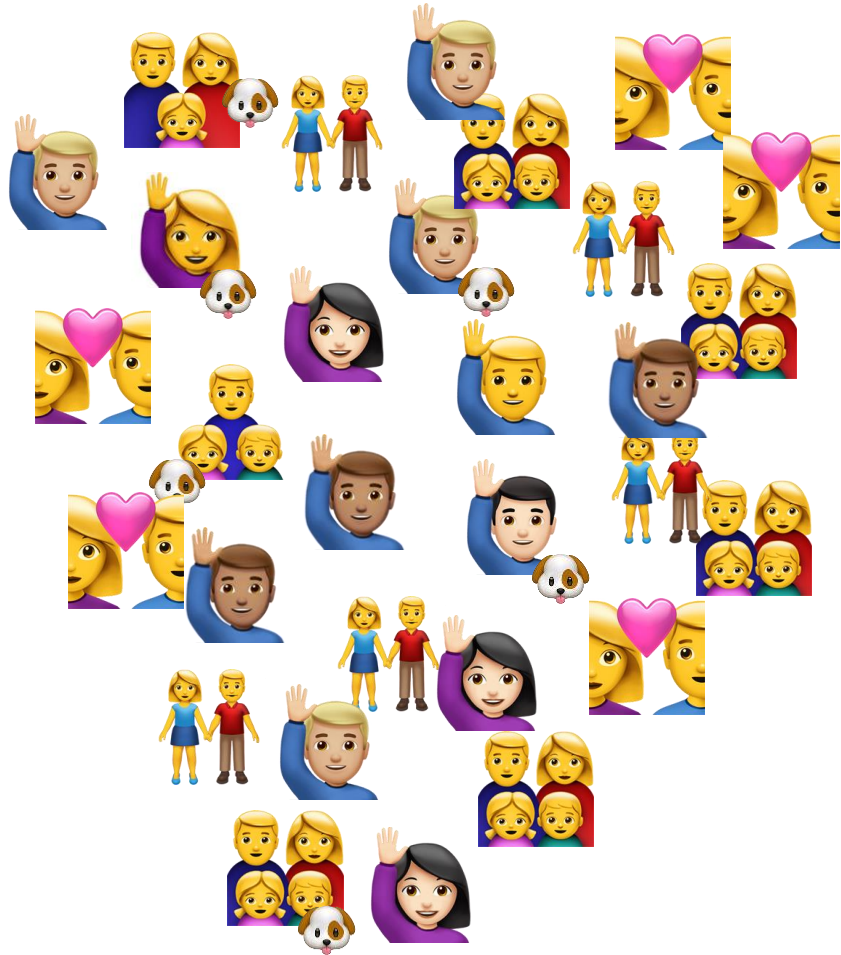
# Personalize Search Results

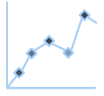
Using Unsupervised ML





# UNDERLYING CHALLENGE



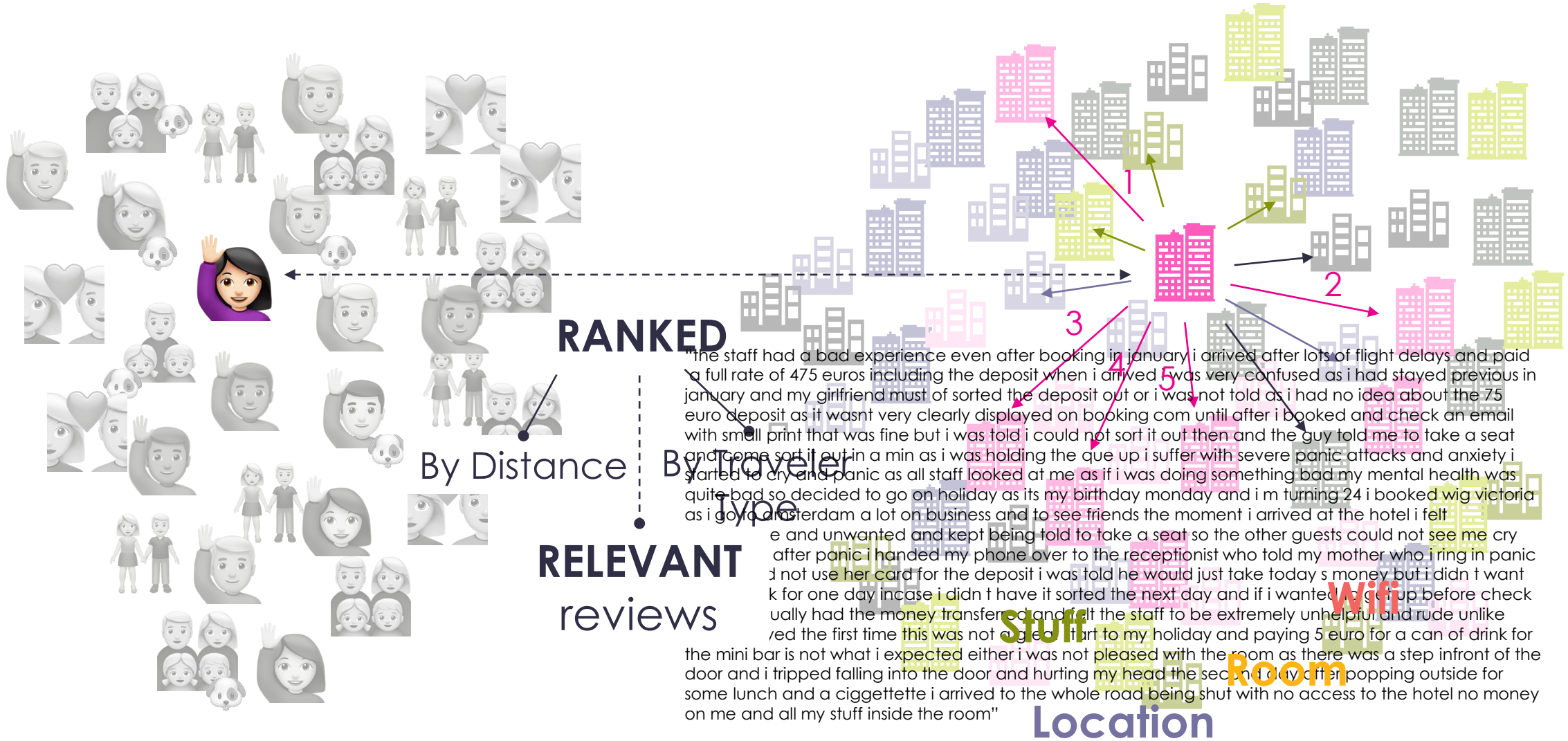


## METHODS

## CLUSTERING

## NLP

## ALGORITHM





# CLUSTERING. HOW ONE SEARCH FOR AN HOTEL?

Automatically assign location tag to the hotel

The image displays two screenshots of hotel listings from a travel website, with various annotations highlighting specific features and user feedback.

**Left Screenshot: DoubleTree by Hilton London – Docklands Riverside**

- Header:** "DoubleTree by Hilton London – Docklands Riverside" with a "Reserve" button and a "Couple friendly" icon.
- Address:** "265 Rotherhithe Street, Southwark, London, SE16 5HW, United Kingdom".
- Image:** A photo of a modern bar with a curved stainless steel counter and brick walls.
- Annotations:**
  - A yellow box labeled "Great location - show map" points to the address.
  - A yellow box labeled "Couples like what this property has to offer" points to the "Couple friendly" icon.
  - A white box labeled "beautiful views b comfortable room" is overlaid on the image.
  - A white box labeled "Staff" is overlaid on the image.

**Right Screenshot: Dorsett City London**

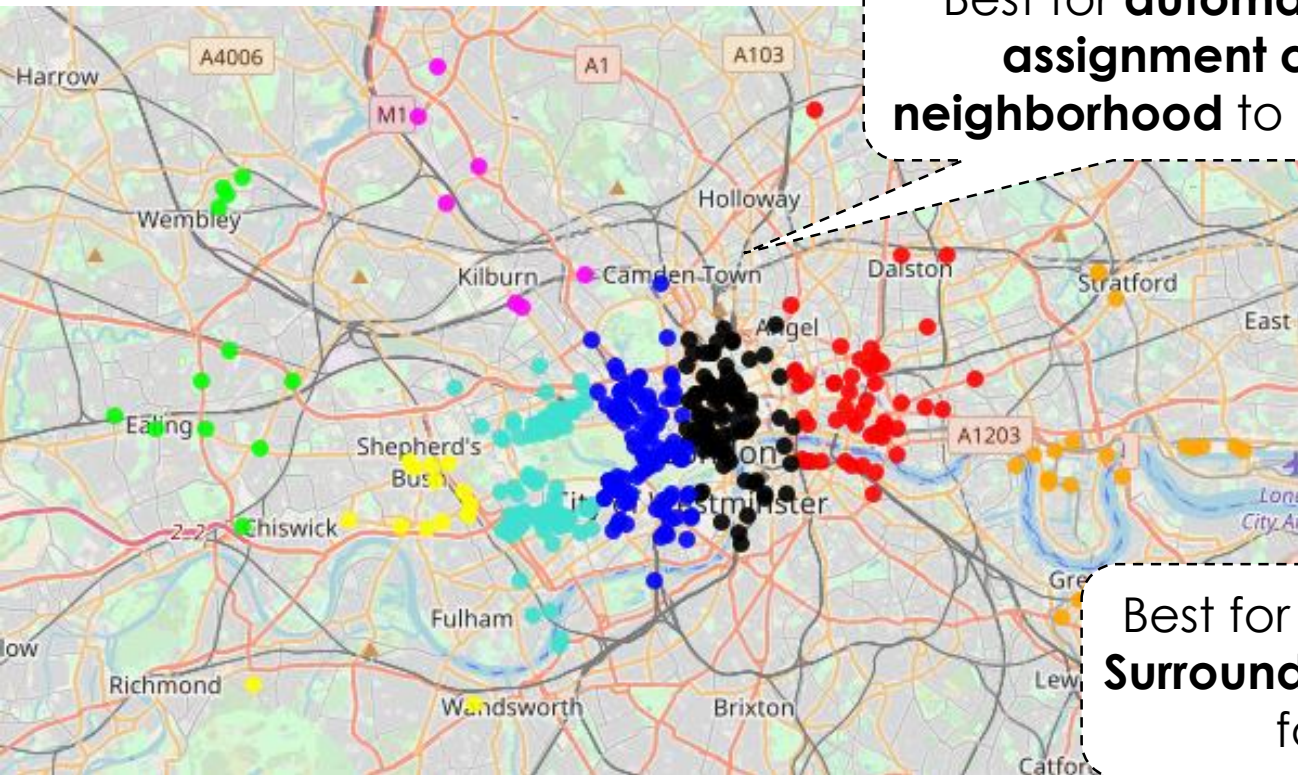
- Header:** "Dorsett City London" with a "Reserve" button and a "Couple friendly" icon.
- Address:** "9 Aldgate High Street, City of London, London, EC3N 1AH, United Kingdom".
- Image:** A photo of a modern dining area with red chairs and wooden tables.
- Annotations:**
  - A yellow box labeled "Excellent location - show map" points to the address.
  - A white box labeled "Superb 9.1" with "1,051 reviews" and a "8.8" score is overlaid on the image.
  - A yellow box labeled "Score from 402 business travellers" points to the "8.8" score.
  - A white box labeled "Great bed, superb bathroom, love the opaque window in the shower." is overlaid on the image.
  - A white box labeled "Excellent location!" with a "9.2" score is overlaid on the image.





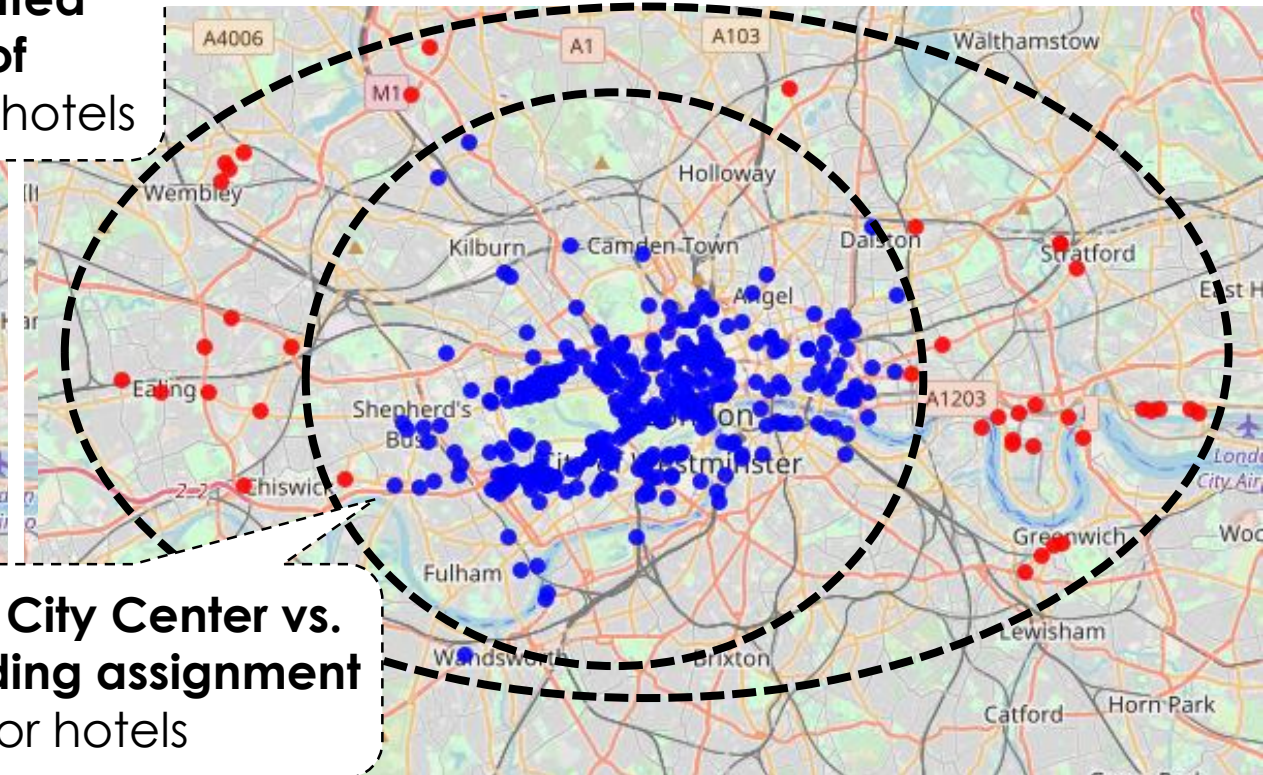
# CLUSTERING. **METHODS.** COMPARISON

## KMEANS



Best for **automated assignment of neighborhood** to hotels

## DBSCAN



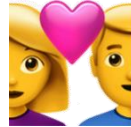
Best for **City Center vs. Surrounding assignment** for hotels

# CLUSTERING. **BUSINESS CHALLENGE**

Personalize hotel search results / recommendations



Yanir

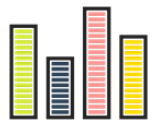


Kobi



Noa





# CLUSTERING. METHODS

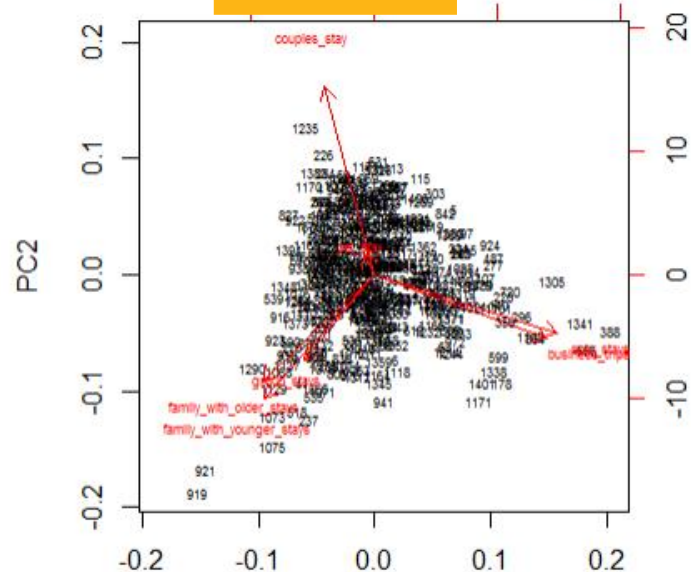
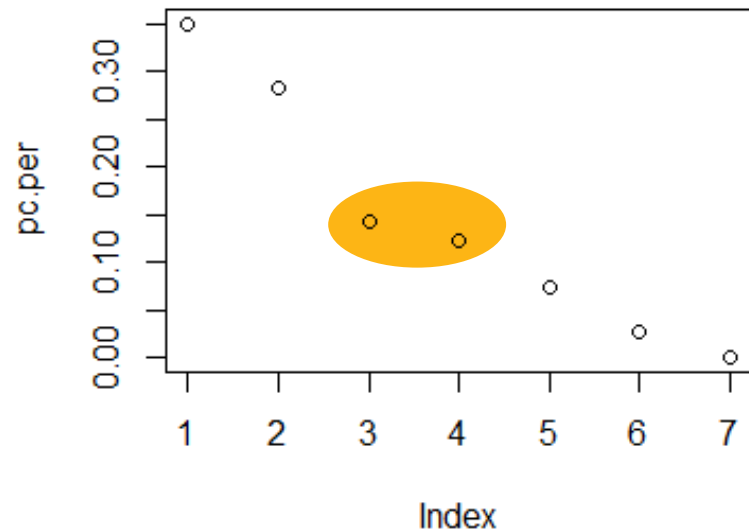
## Dimensionality Reduction - PCA

standard deviations (1, ..., p=7):

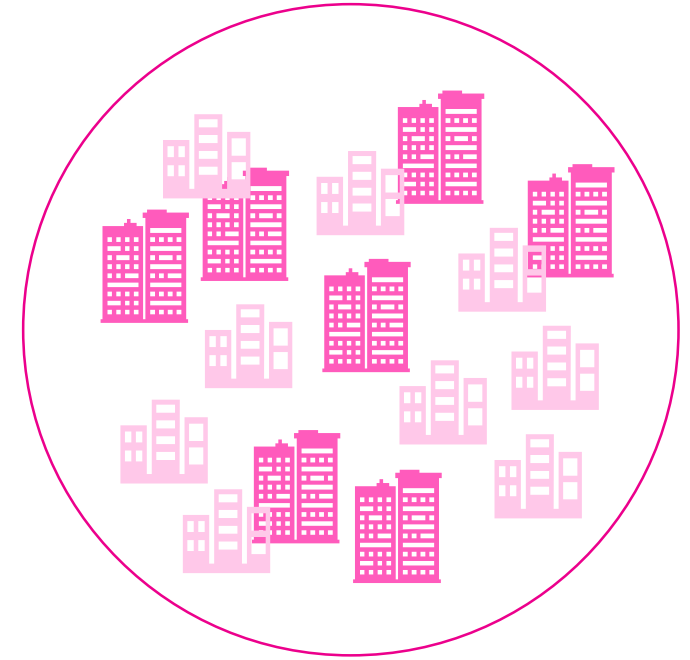
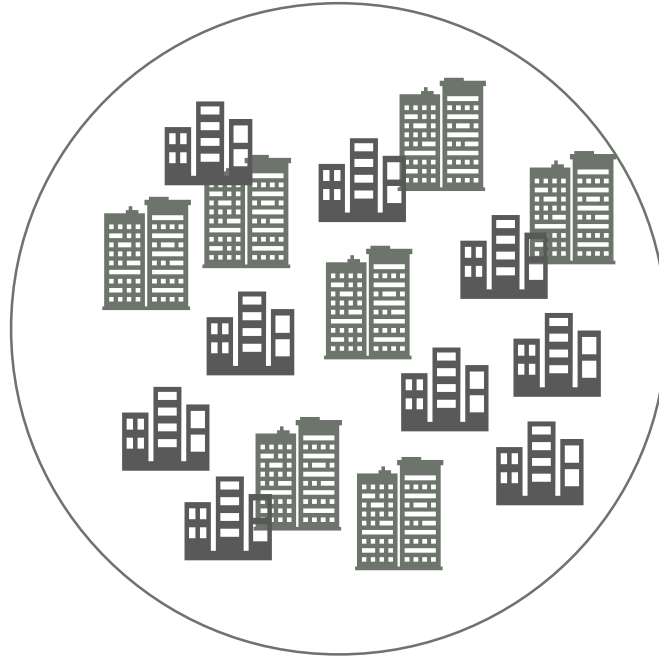
[1] 1.56640718 1.40685881 0.99908959 0.92565552 0.72493386 0.43140421 0.02144696

Rotation (n x k) = (7 x 7):

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
🕶️ business_trips	0.55977527	-0.22907968	-0.06405424	0.08461212	0.05777924	0.78712880	0.001347057
💕 couples_stay	-0.16159414	0.67840677	0.05334131	0.03222923	0.08345737	0.30601295	0.639654996
👨👩👦 pet_stay	-0.03539911	0.07769089	-0.98933010	0.06462701	0.08737650	-0.04608224	-0.001449973
👨👩👦 family_with_older_stays	-0.36454270	-0.38116160	0.02728986	-0.27986336	0.77503813	0.12345083	0.163804791
👨👩👦 family_with_younger_stays	-0.35503287	-0.44382457	-0.10640842	-0.31828719	-0.61029915	0.19299941	0.393227788
🕶️ solo_stays	0.59072207	-0.21226768	-0.01386827	-0.08819196	0.09234030	-0.48132550	0.598179531
group_stays	-0.22670158	-0.30647822	0.04492096	0.89454869	0.01964406	-0.02230787	0.227065444

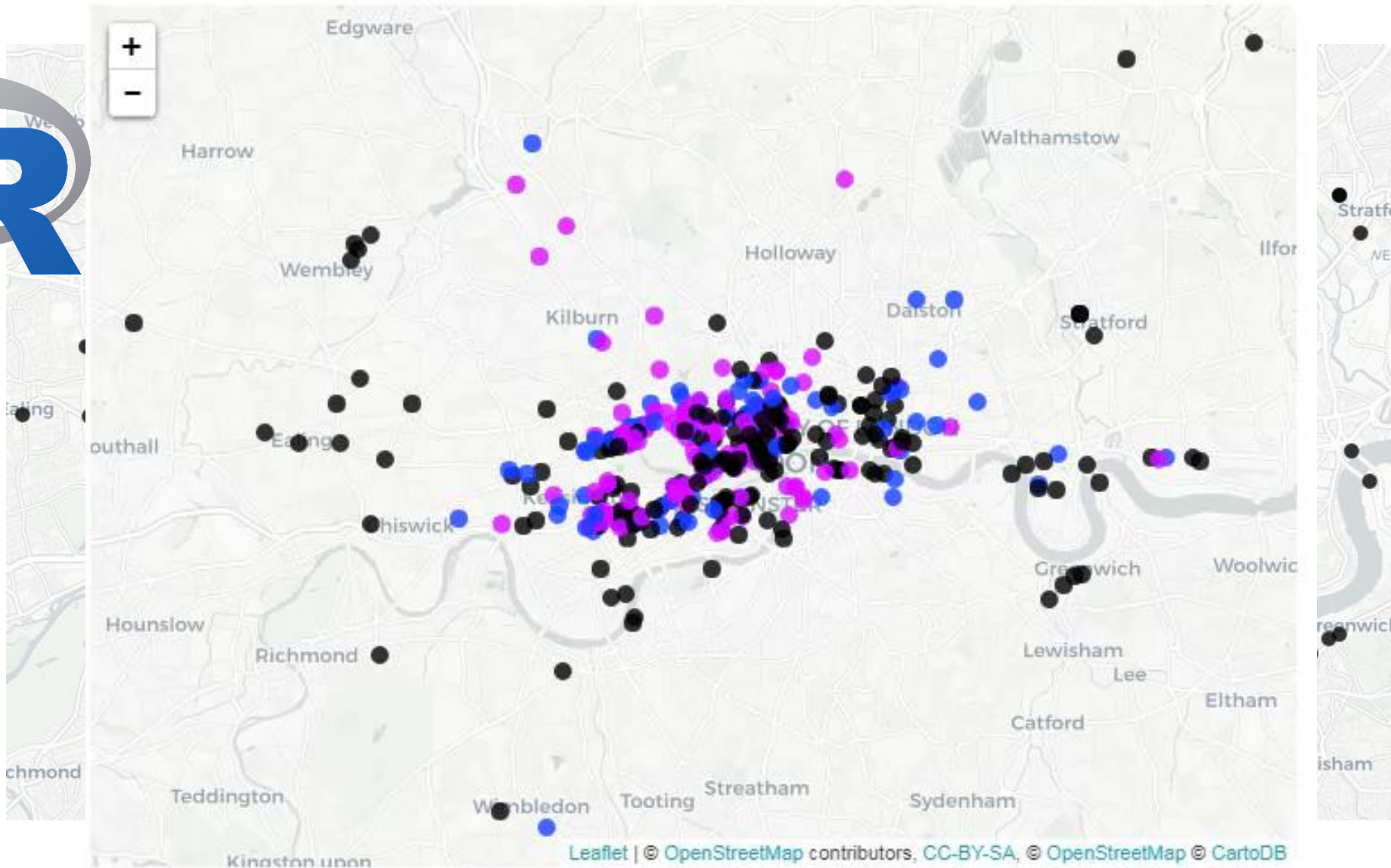


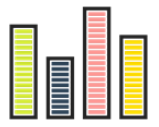
# CLUSTERING. **METHOD**





# CLUSTERING. ON MAP





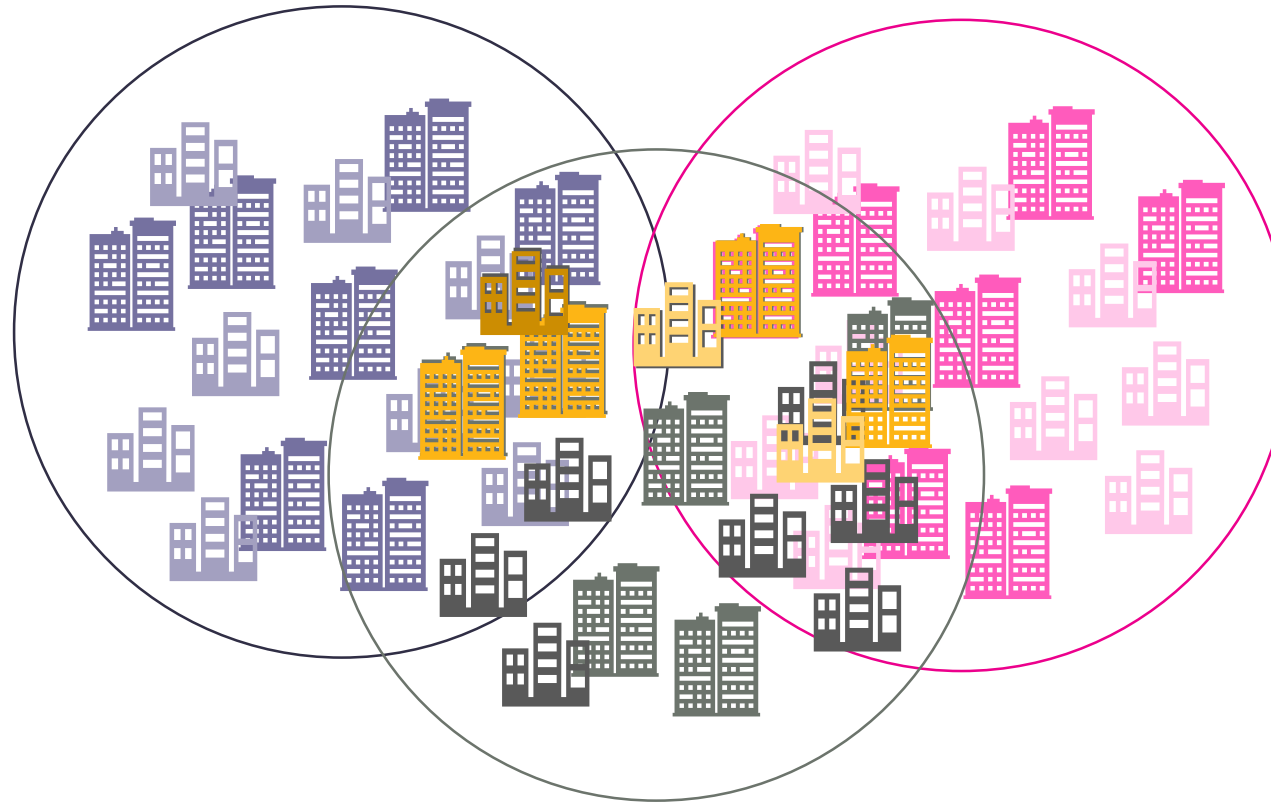
## CLUSTERING. **BUSINESS OBJECTIVES MET**

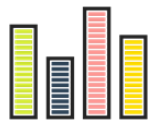
Which business problems / predictions we could fulfill?

1. Machine Learning model (clustering) to **personalize search results**
2. **Prediction model for Segmentation** correlated to geo-location
3. **Recommendation engine for hotels by proximity**

# CLUSTERING. **BUSINESS SOLUTION – REVIEW MODEL**

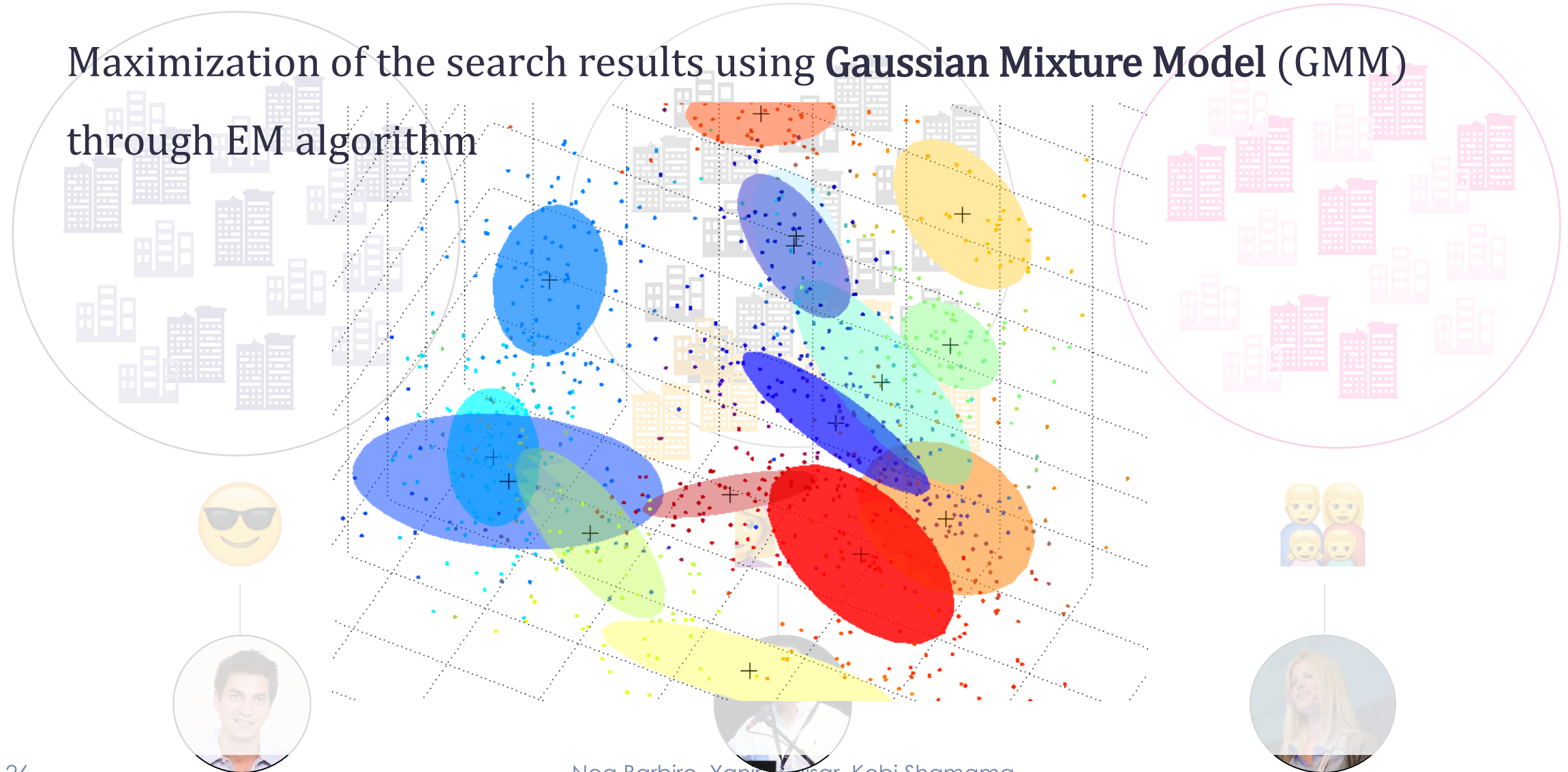
What are we missing with the previous algorithm?





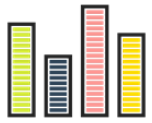
# CLUSTERING. BUSINESS SOLUTION – REVIEW MODEL

Maximization of the search results using **Gaussian Mixture Model (GMM)**  
through EM algorithm









NLP. BAG OF WORDS

ITY GIVER SMILE  
JOB HELP

HAPPY IDENTIFY LOCATION SMILE  
RECRUIT GIVER PHILANTHROPY CARE HELP



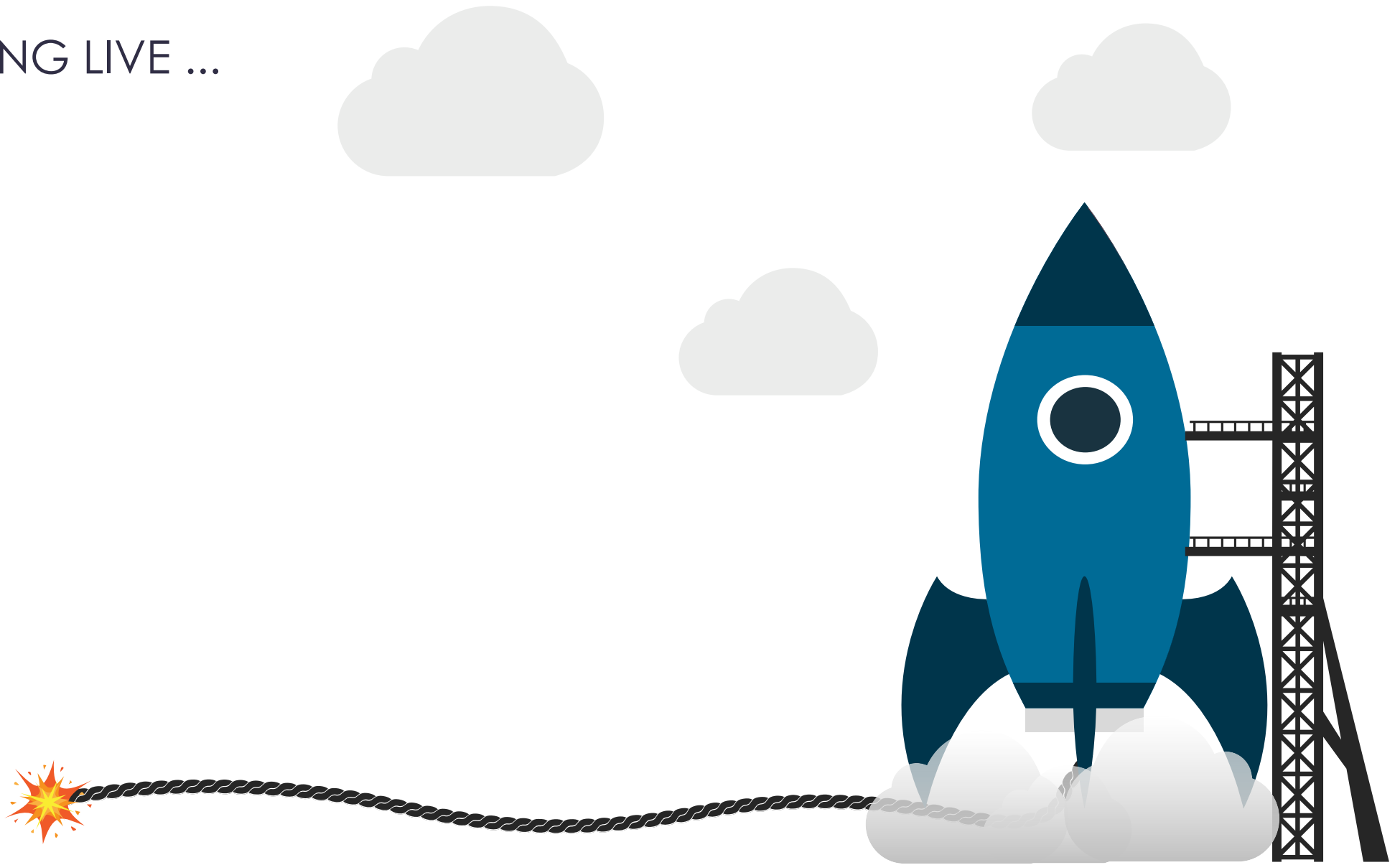
LOCATION  
CARE  
WALKING  
LOVE  
BREAKFAST  
STAY  
NICE  
BED  
VALUABLE  
COMFY  
LONDON  
EVERYTHING

STAFF

WELCOMING  
MODERN  
BATHROOM  
FACILITIES  
LIKE  
SERVICE  
GOOD  
VIEW  
HOTEL  
FRIENDLY  
WILLING  
EXCEL  
GROUP  
EXCELLENT  
ROOM  
NICE  
STATION  
EXCELLENT  
SHOWER  
TUBE  
SMILE  
FOOD  
SHOWER  
GO  
JOIN  
WELL  
TIM  
ENERGY  
FOUNDATION  
IDENTIFY  
GIFTING  
MODERN  
SUCCESS  
SACRIFICE

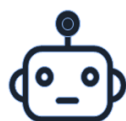
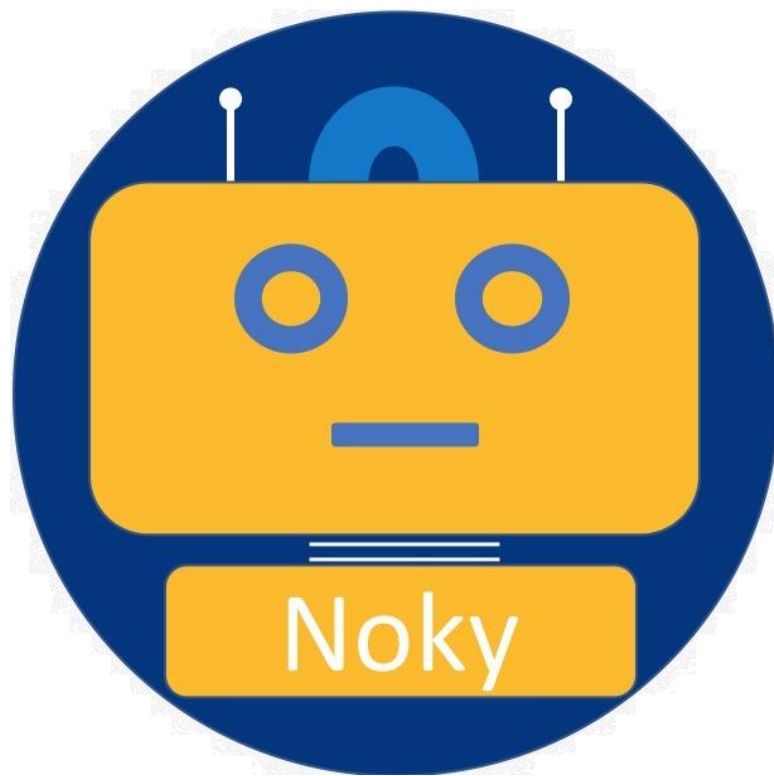


 GOING LIVE ...



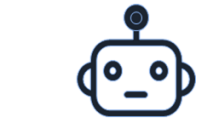
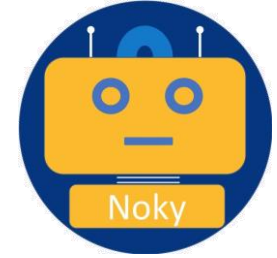


INTO ACTION...



MEET NOKY (Noa-Kobi-Yanir)

# BUSINESS APPLICATION. CHATBOT



User starts interaction

(greeting)  
intent extraction  
question: 'How can I help  
you today?'

Enrichment for traveler profile:

Nationality  
Gender  
Destination  
Trip Type

NLP to detect  
Intent and Context

(understanding amenities)  
question: 'Are you interested  
with something in particular?'

Use clustered data,  
ranked by the different  
amenities identified in  
the NLP analysis

Integration to Open Source API

Recommended  
Hotels



Clustered data to predict  
and optimize questions  
asked by bot

No

Yes



Feedback loop to  
optimize results

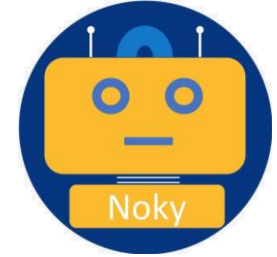
Continue for booking  
at **Booking.com**







# BUSINESS APPLICATION. CHATBOT



So, how can y



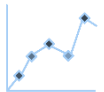
When

**Traveling** from Israel

ould I book at **London, UK**?

ive **Wifi** working ..





## SUMMARIZING



### INTERACTIVE CHATBOT

Open source API for interactive personalized search results.



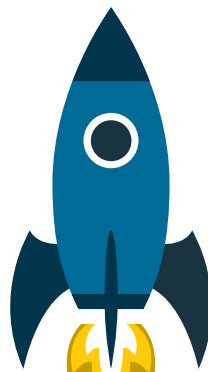
### CLUSTERING

Clustering methods including KMEANS and DBSCAN  
Dimensionality reduction using CPA



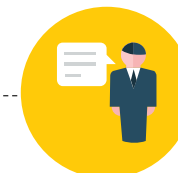
### EXPLORATORY DATA ANALYSIS

Feature engineering & data enrichment



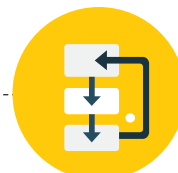
### NATURAL LANGUAGE PROCESSING

NLP using uni-gram, tokenization, TF-IDF  
Analysis LDA topic clustering



### DATA CORRELATIONS

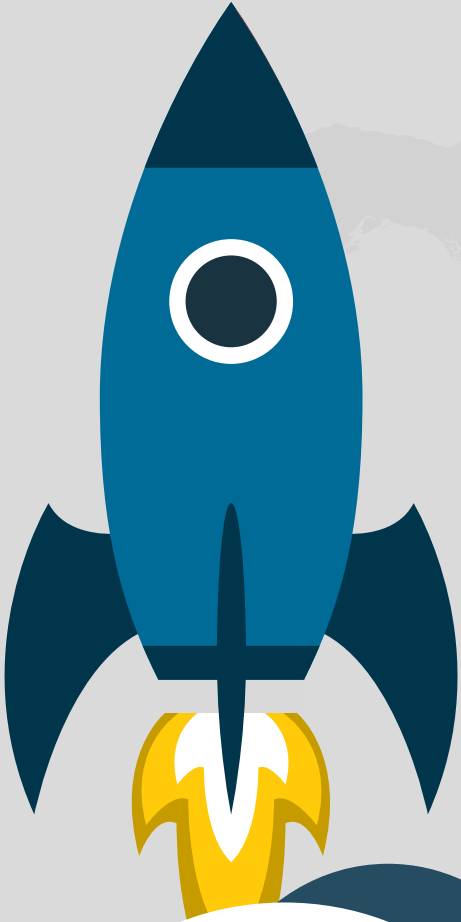
Logistic and linear models



# WHAT DID WE HAVE?

# THE TEAM

PERSONALIZE YOUR HOTEL BOOKING USING MACHINE LEARNING



**Yanir Calisar**

Head of Intelligence  
& Partnerships



**Noa Barbiro**

Head of Product  
& Strategy



**Kobi Shamama**

Head of Analytics  
& Machine Learning

A close-up photograph of a brass key inserted into a lock on a white door. The key has a distinctive octagonal head and is attached to a brass chain with a rectangular tag. The background is a blurred interior space with warm, orange-toned lighting. A semi-transparent white horizontal band across the middle of the image contains the text "ENJOY YOUR STAY" in a large, black, sans-serif font.

ENJOY YOUR STAY