# Personalize Hotel Booking

## Analyzing Hotel Reviews over Booking.com

Noa Barbiro
Yanir Calisar
Kobi Shamama

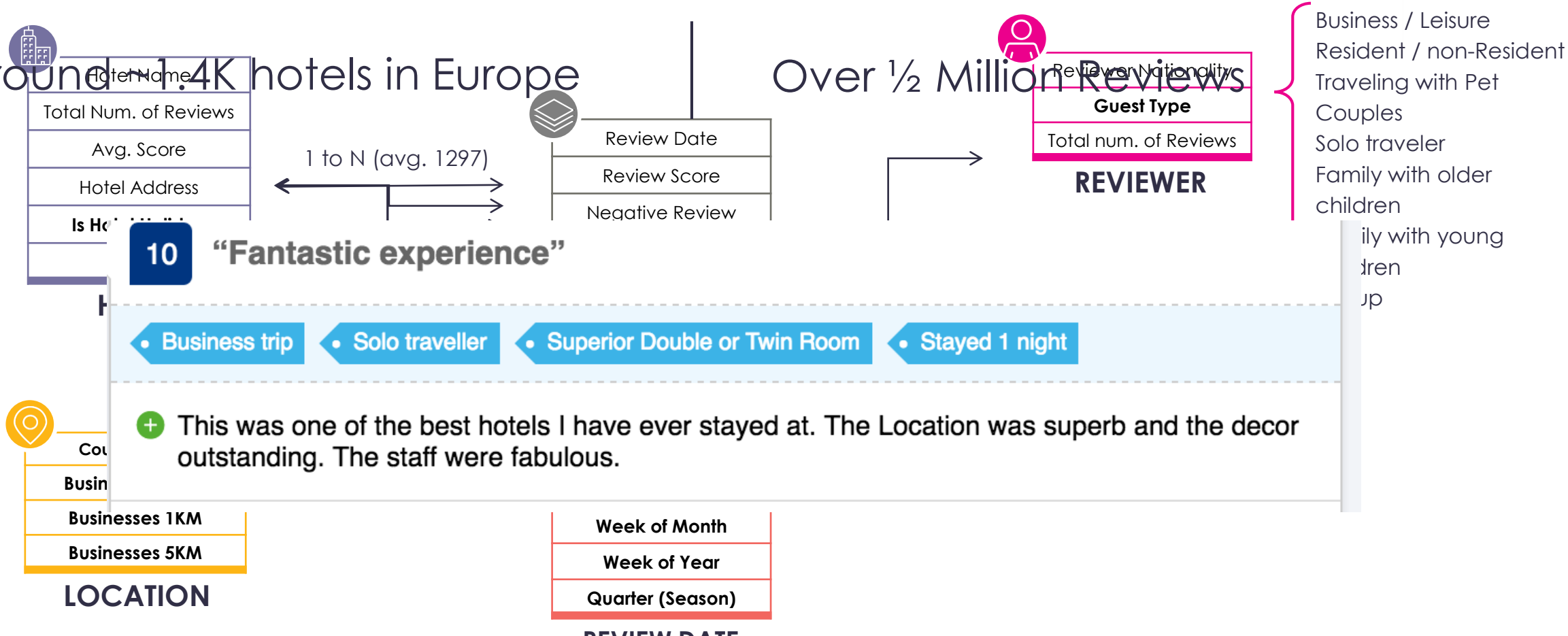
kaggle Booking.com yelp Google Maps

# SO, WHAT DO WE HAVE HERE?

Booking.com yelp Google Maps

## Around 1.4K hotels in Europe

## Over ½ Million Reviews






Noa Barbiro, Yanir Calisar, Kobi Shamama

## Contributing Back to the Community

kaggle Search kaggle Competitions Datasets Kernels Discussion Jobs ••• Sign In

Published Dataset

# Enriched Hotel Reviews Dataset

1

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

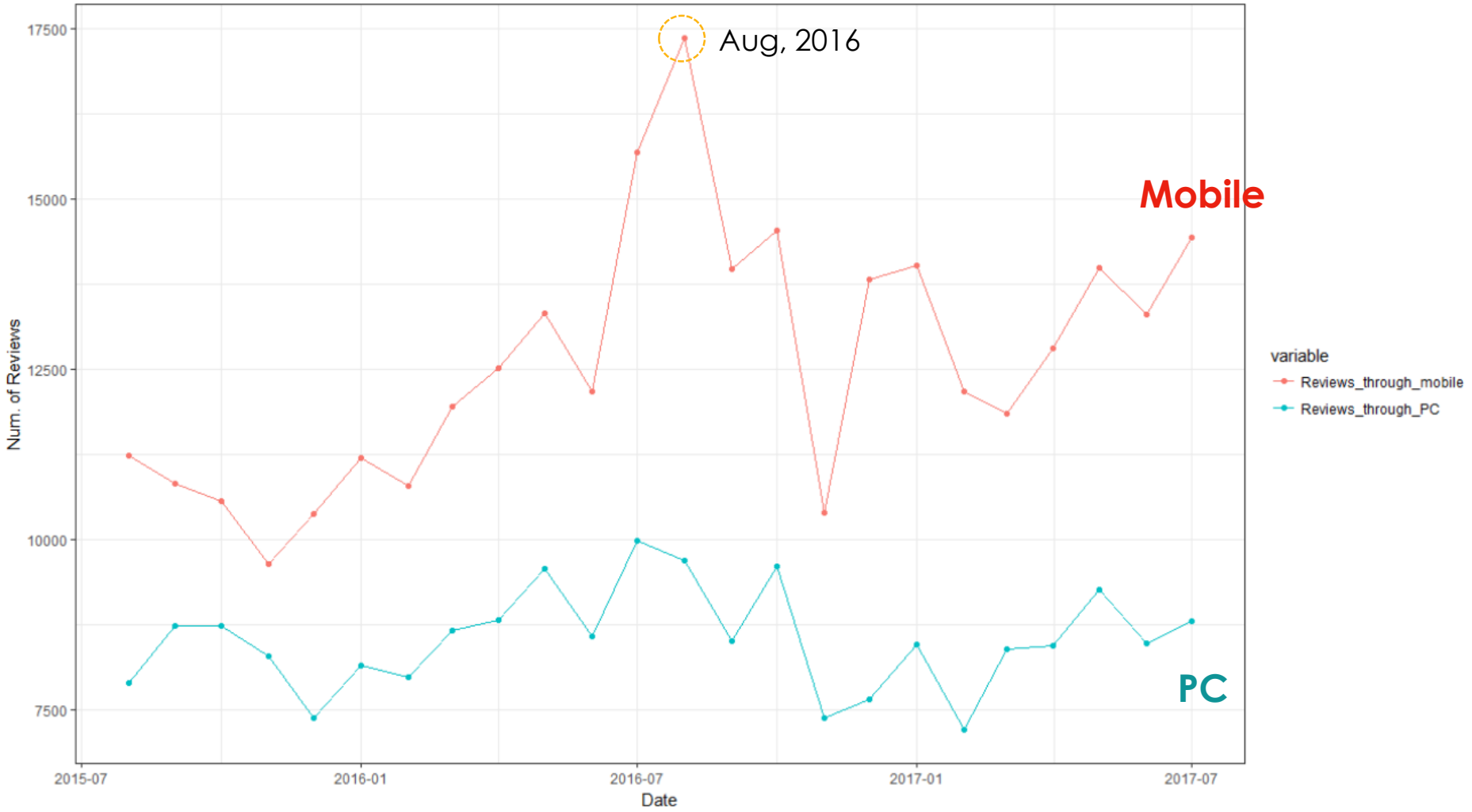
Preview (first 100 rows) Column Metadata Column Metrics

| 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 |

# SOME COOL STATS

# EDA IN ACTION. **FINDINGS**

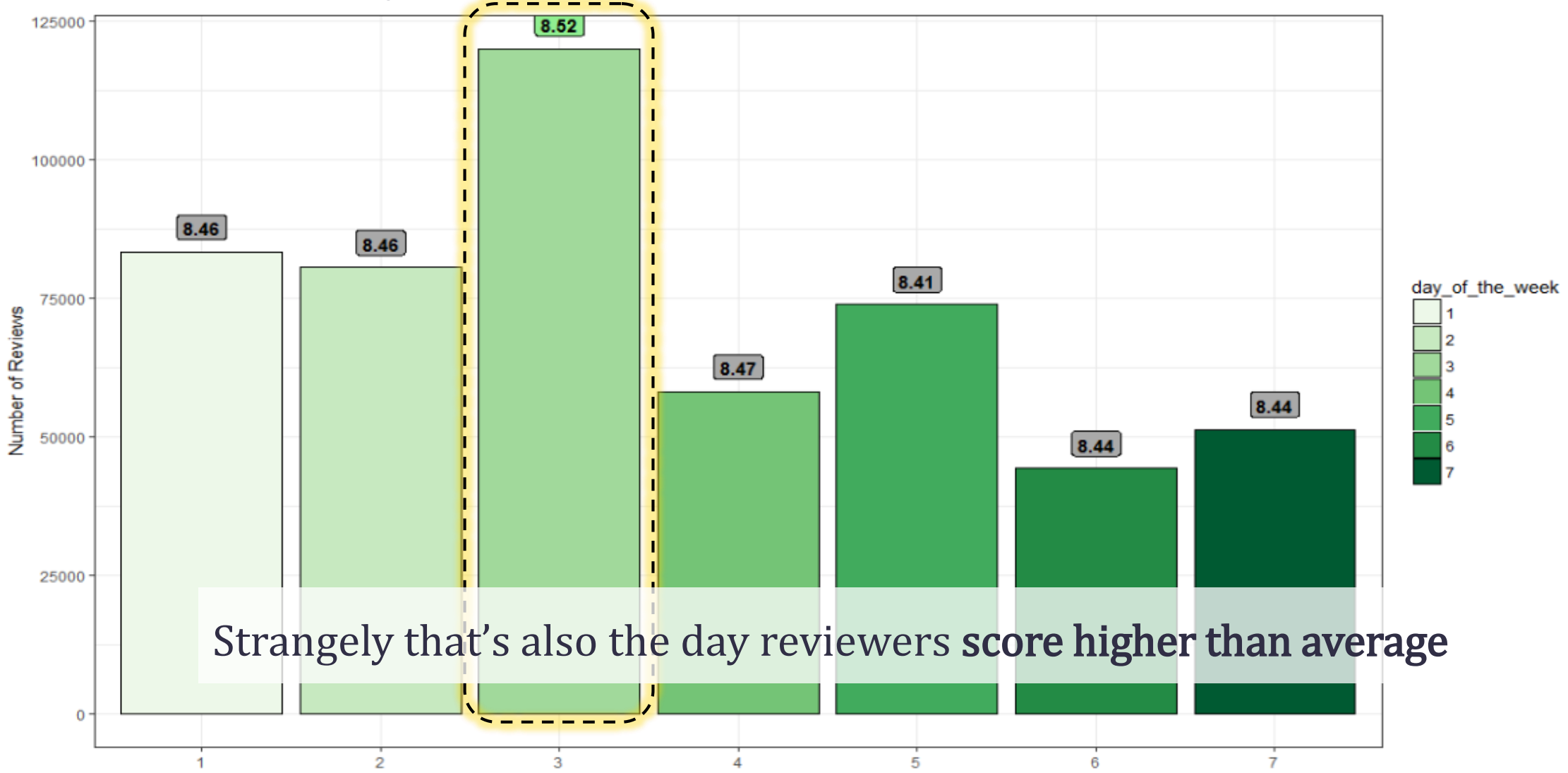
Over time **mobile is used more than pc** to submit reviews, almost ~200% more



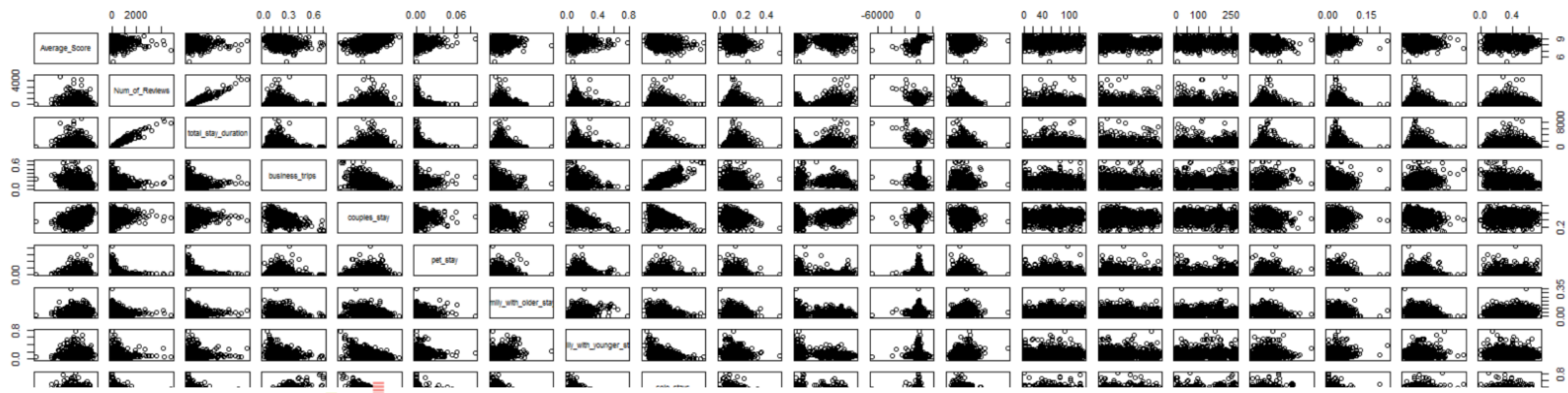
# EDA IN ACTION. **FINDINGS**

TUESDAY

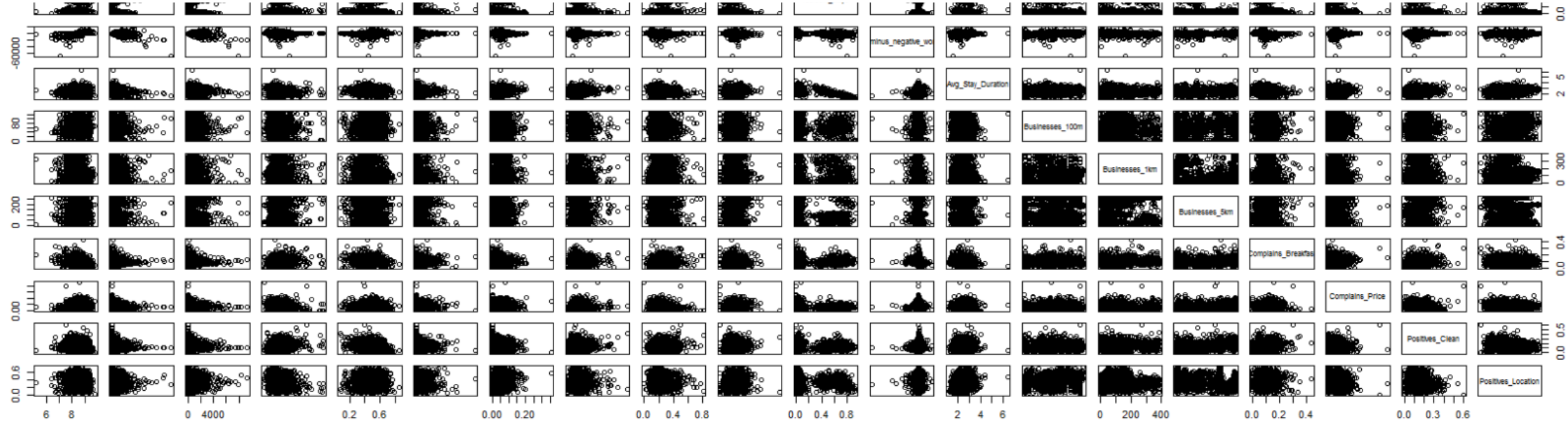
The **most popular day** for submission of a review is Tuesday




Strangely that's also the day reviewers **score higher than average**



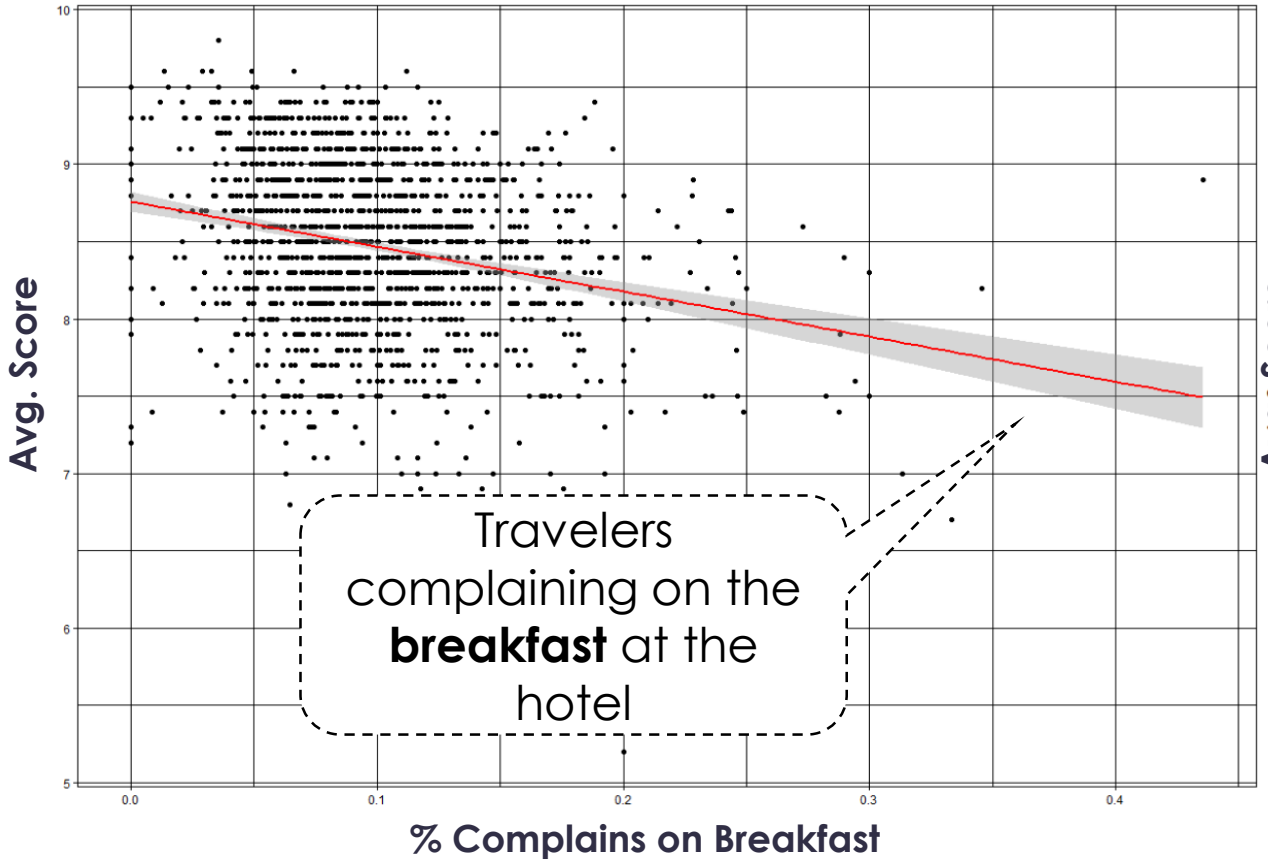

# INTERESTING CORRELATIONS

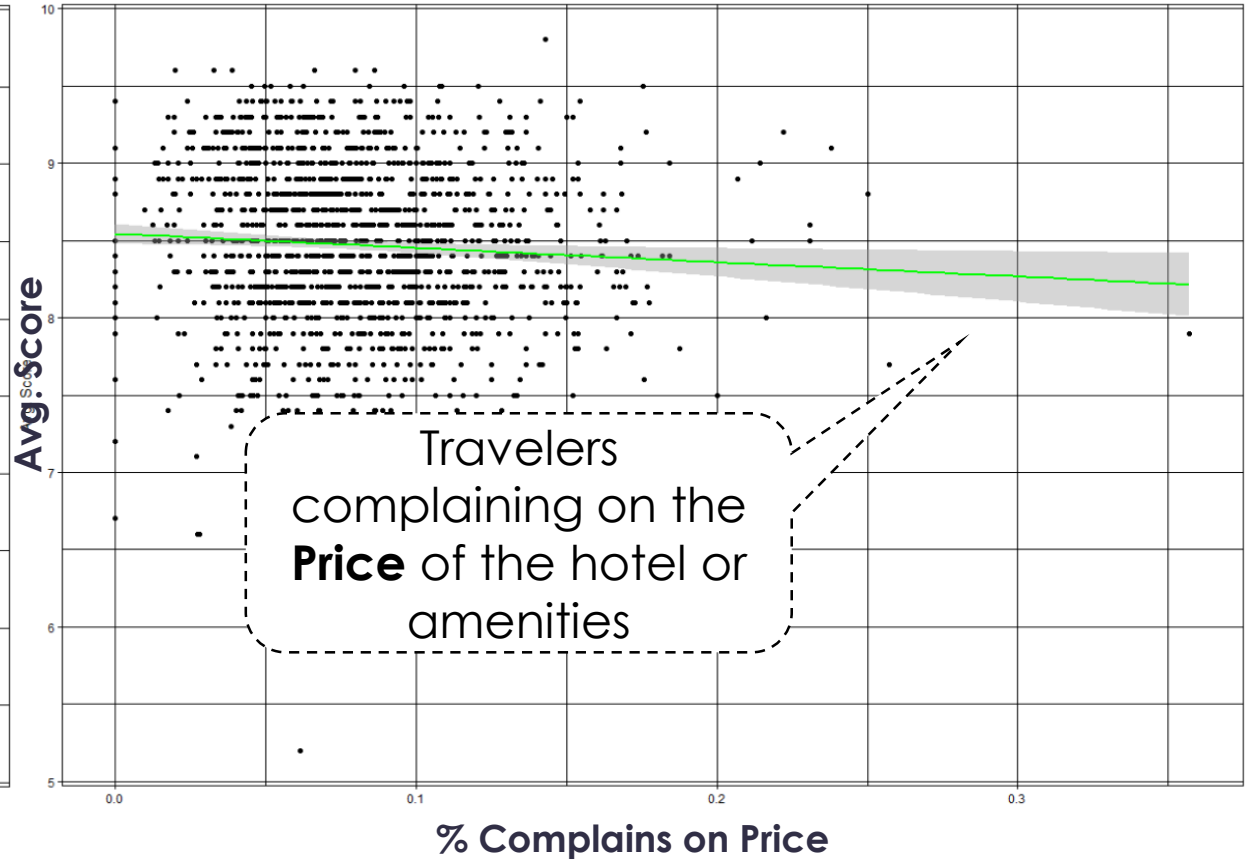


# CORRELATION & REGRESSION. **FINDINGS**

**Bad Breakfast** tends to impact review score

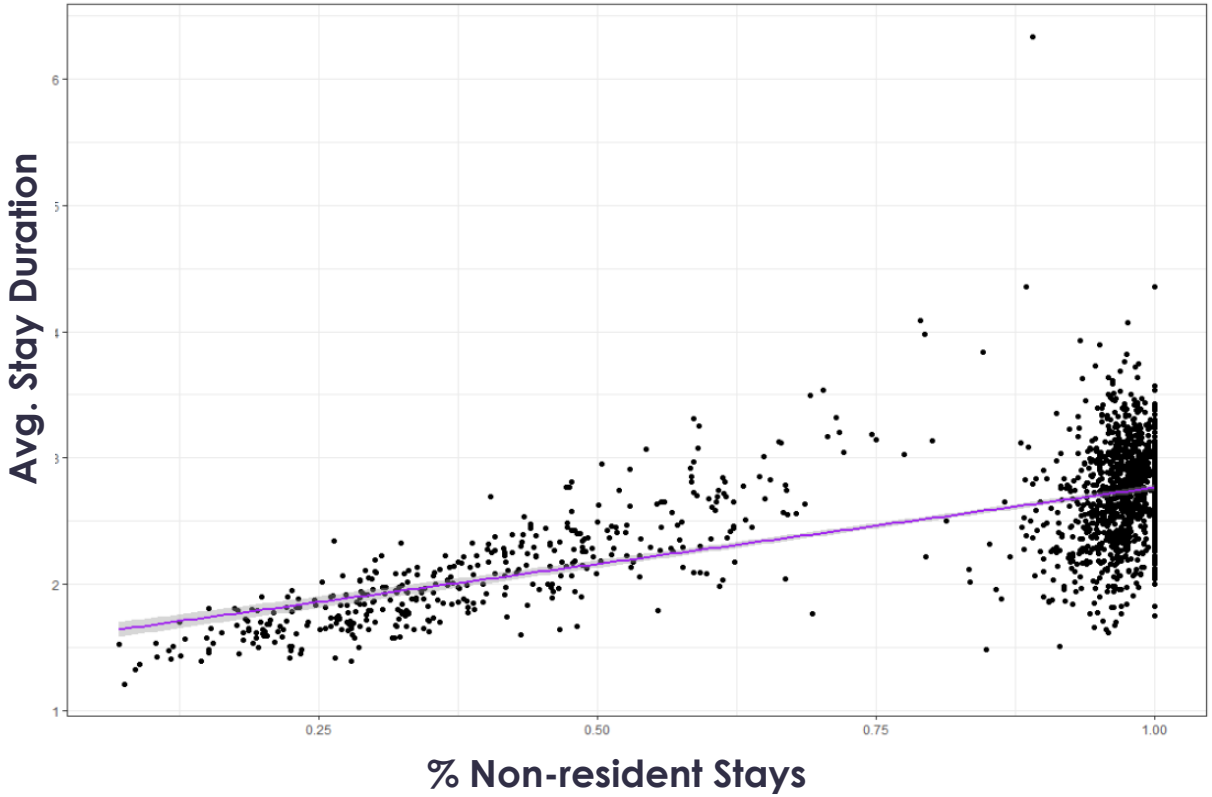
.. But not so much for **high price** !



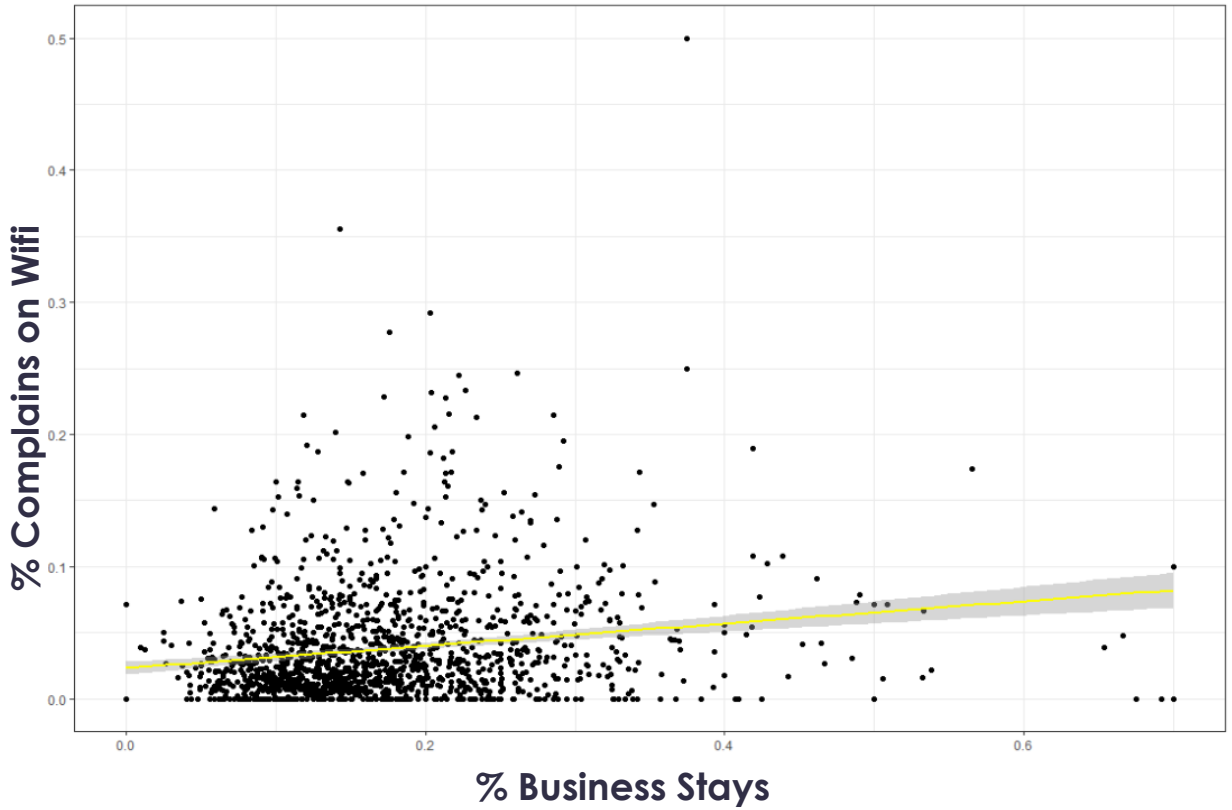



# CORRELATION & REGRESSION. **FINDINGS**

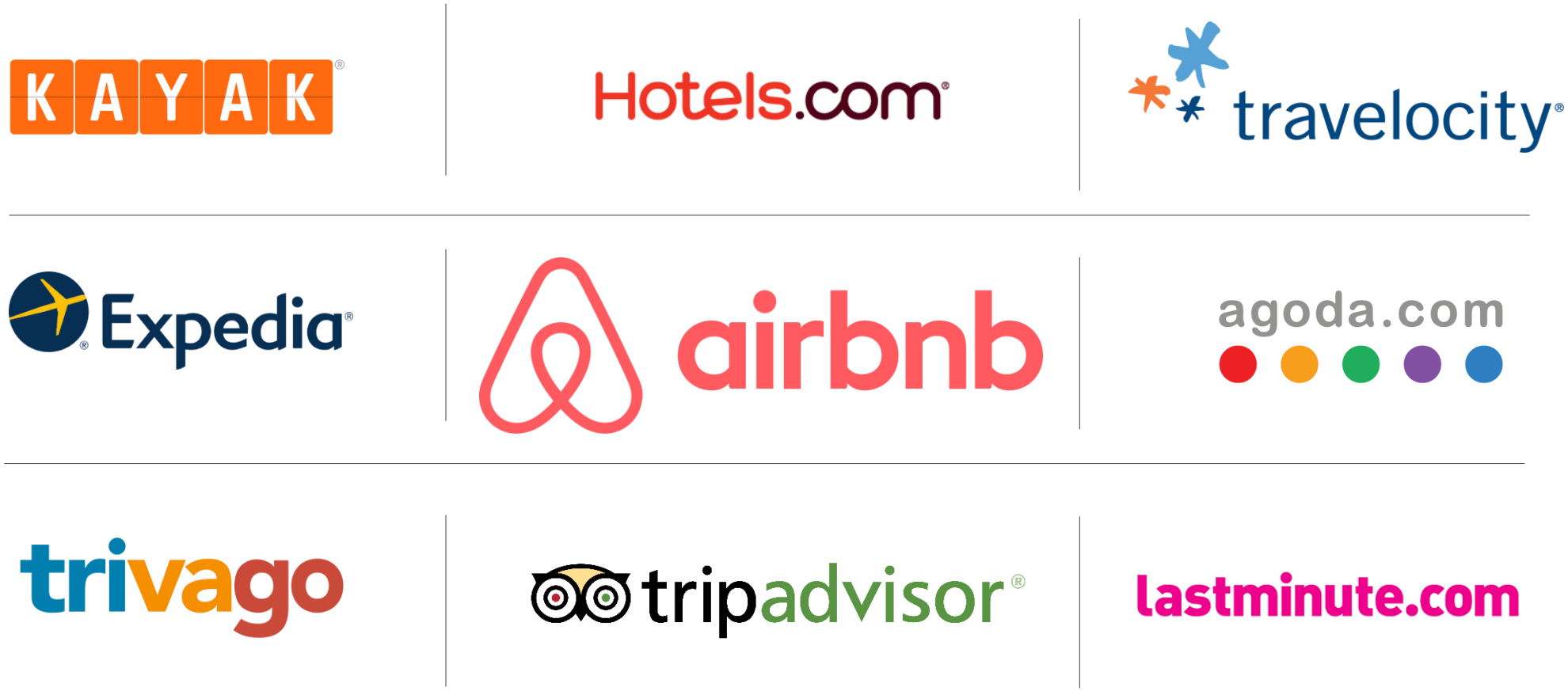
**Non-Resident** Travelers correlate with **longer avg. stay duration**




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



# UNDERSTANDING THE LANDSCAPE



# WHAT IS SO UNIQUE WITH AIRBNB ?

BUSINESS INSIDER

TECH FINANCE POLITICS STRATEGY LIFE ALL

## Machine learning is driving growth at Airbnb

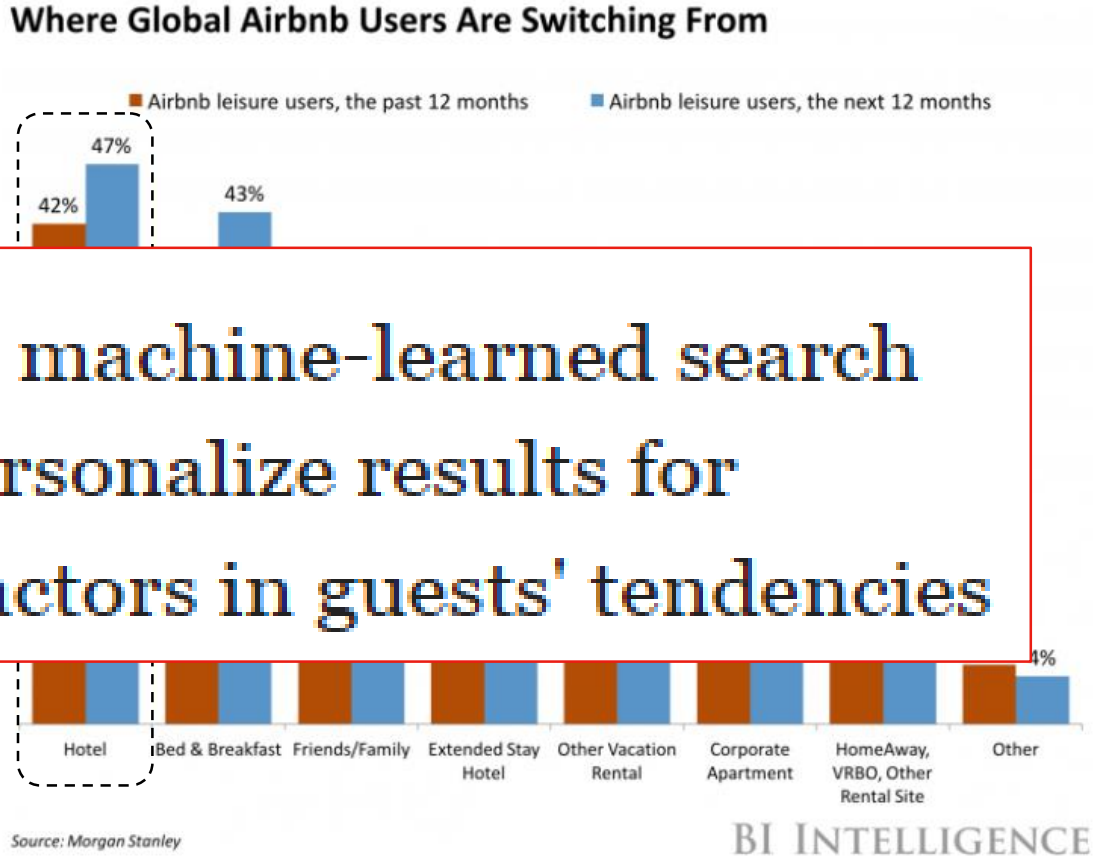
Stephanie Pandolph

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




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

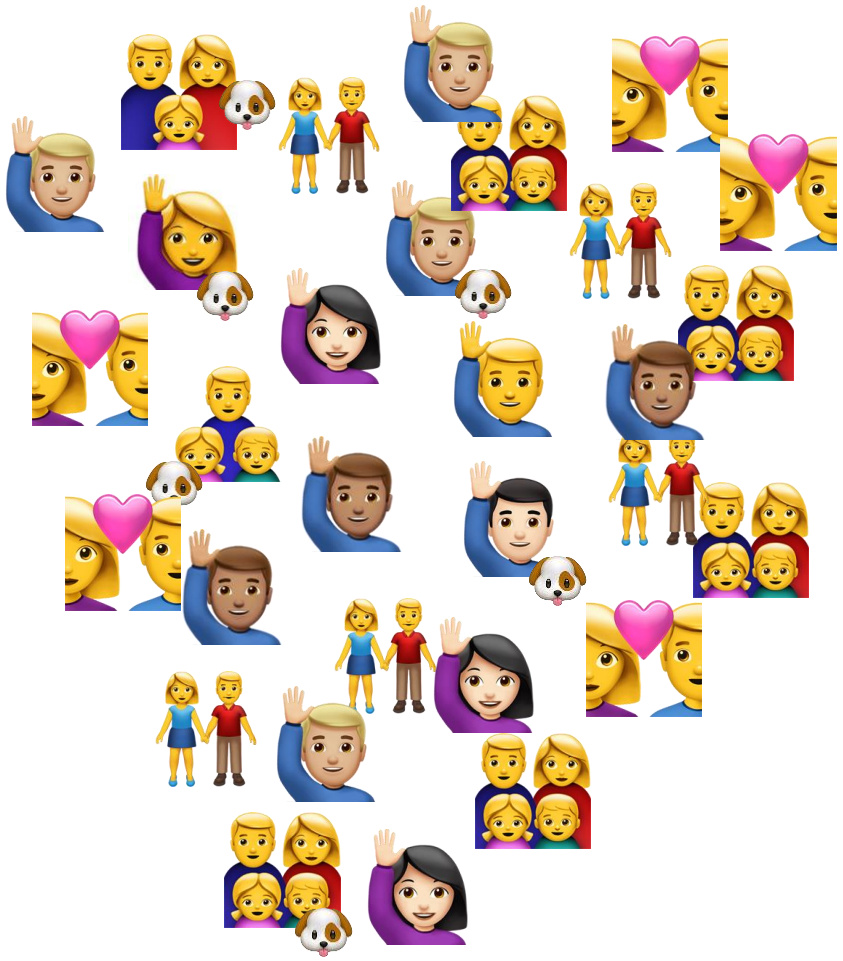
specific types of décor in places they book. The company feeds more than 100 characteristics into the model, which then uses the data to identify patterns and personalize search rankings.

BUSINESS OBJECTIVES

# Personalize Search Results

## Using Unsupervised ML

# UNDERLYING CHALLENGE

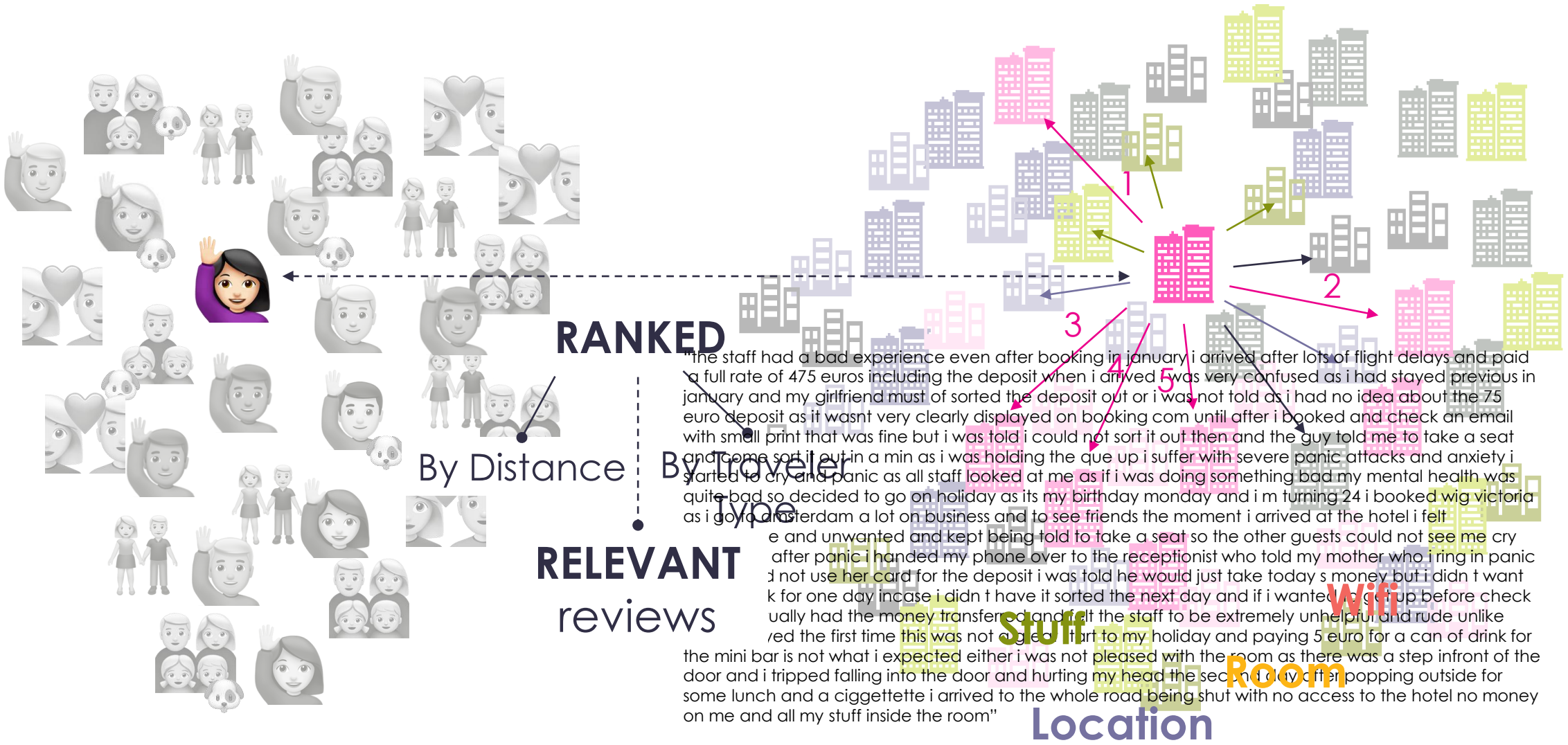


# METHODS

## CLUSTERING

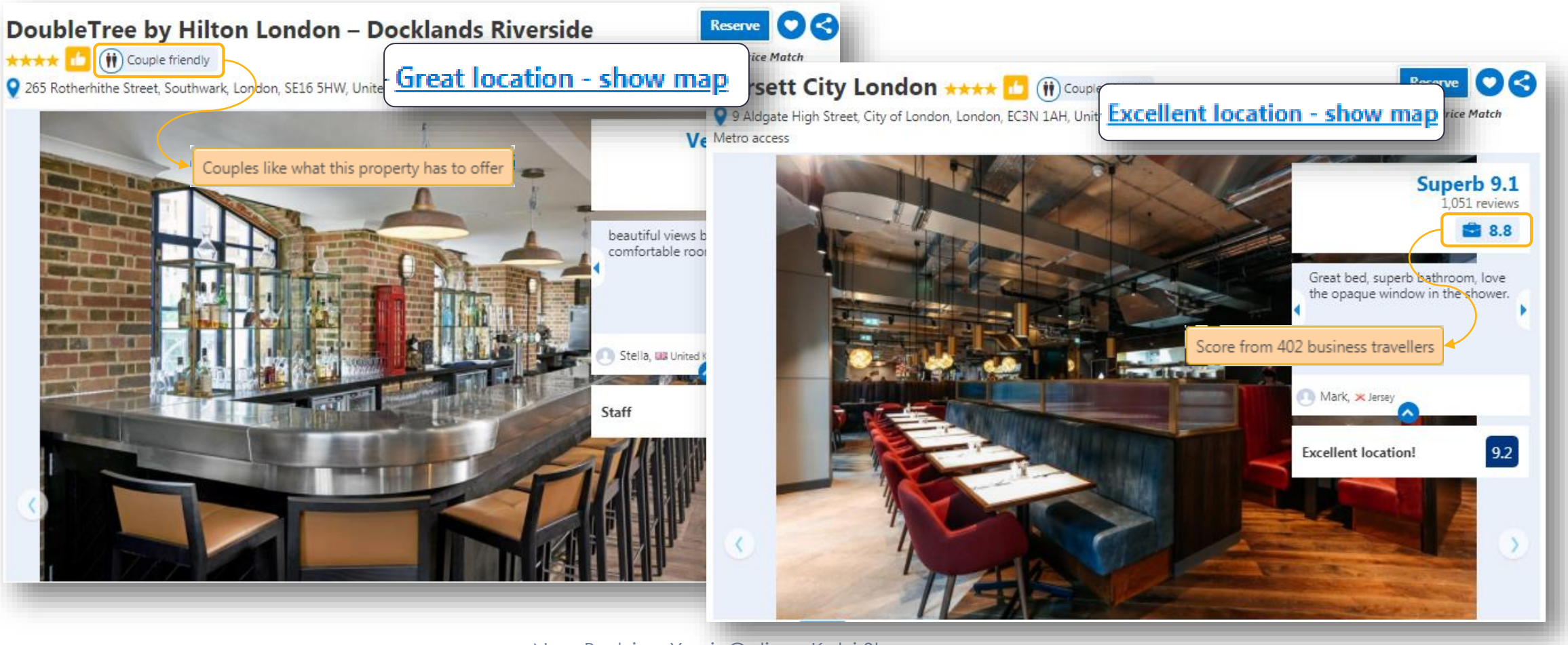
## NLP

## ALGORITHM



# CLUSTERING. **HOW ONE SEARCH FOR AN HOTEL?**

Automatically assign location tag to the hotel



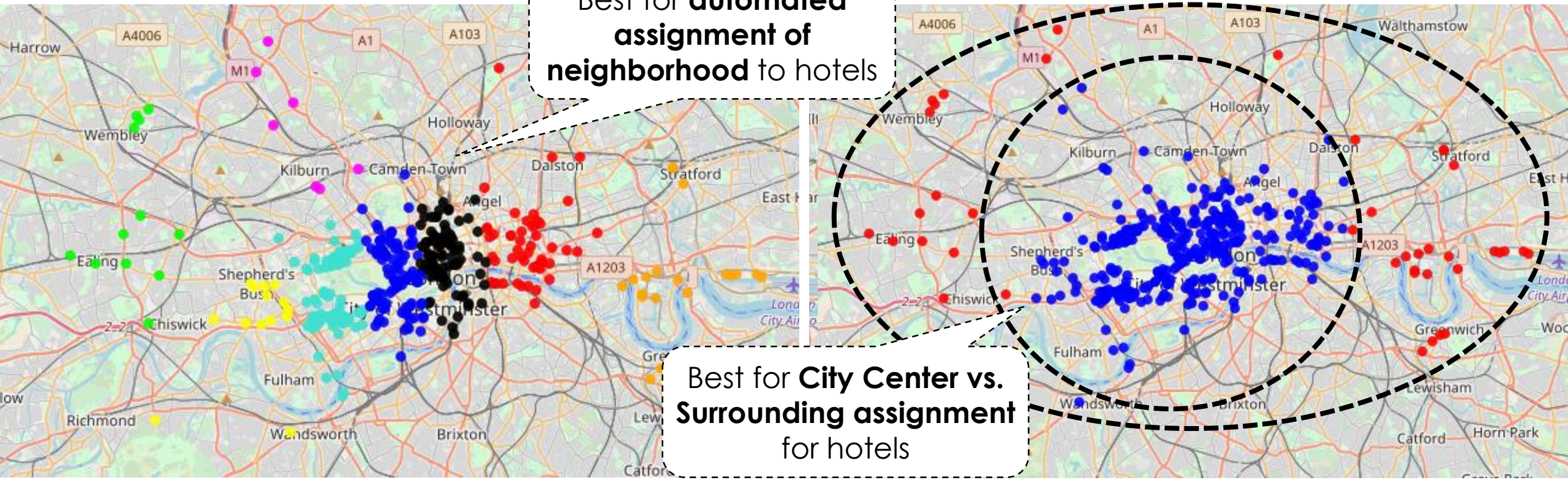
# CLUSTERING. **METHODS.** COMPARISON

## KMEANS

## DBSCAN

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

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



# Personalize hotel search results / recommendations

Yanir

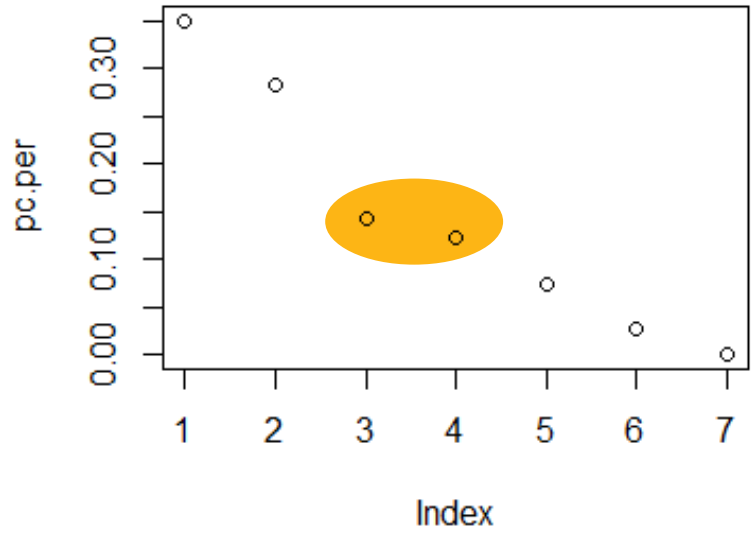
Kobi

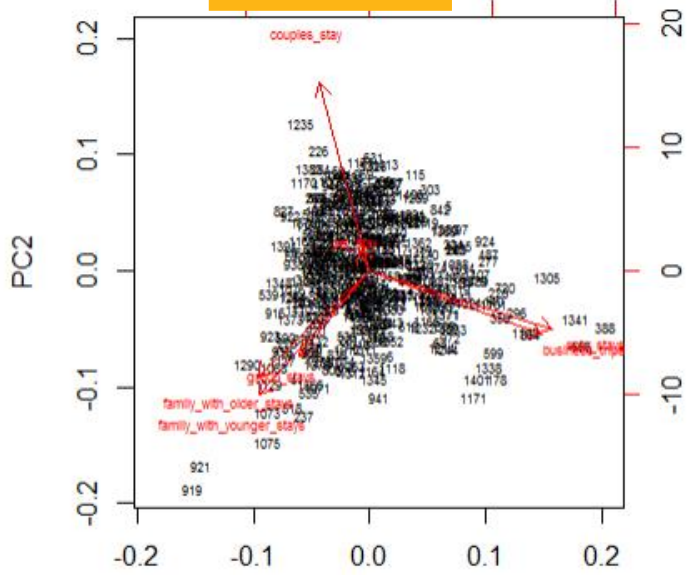
Noa

# Dimensionality Reduction - PCA

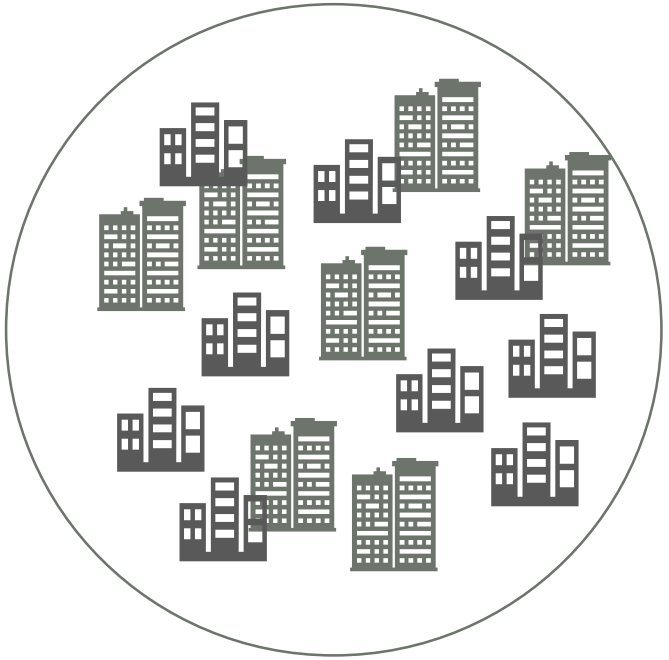
```
Standard deviations (1, .., p=7):
[1] 1.56640718 1.40685881 0.99908959 0.92565552 0.72493386 0.43140421 0.02144696

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
```

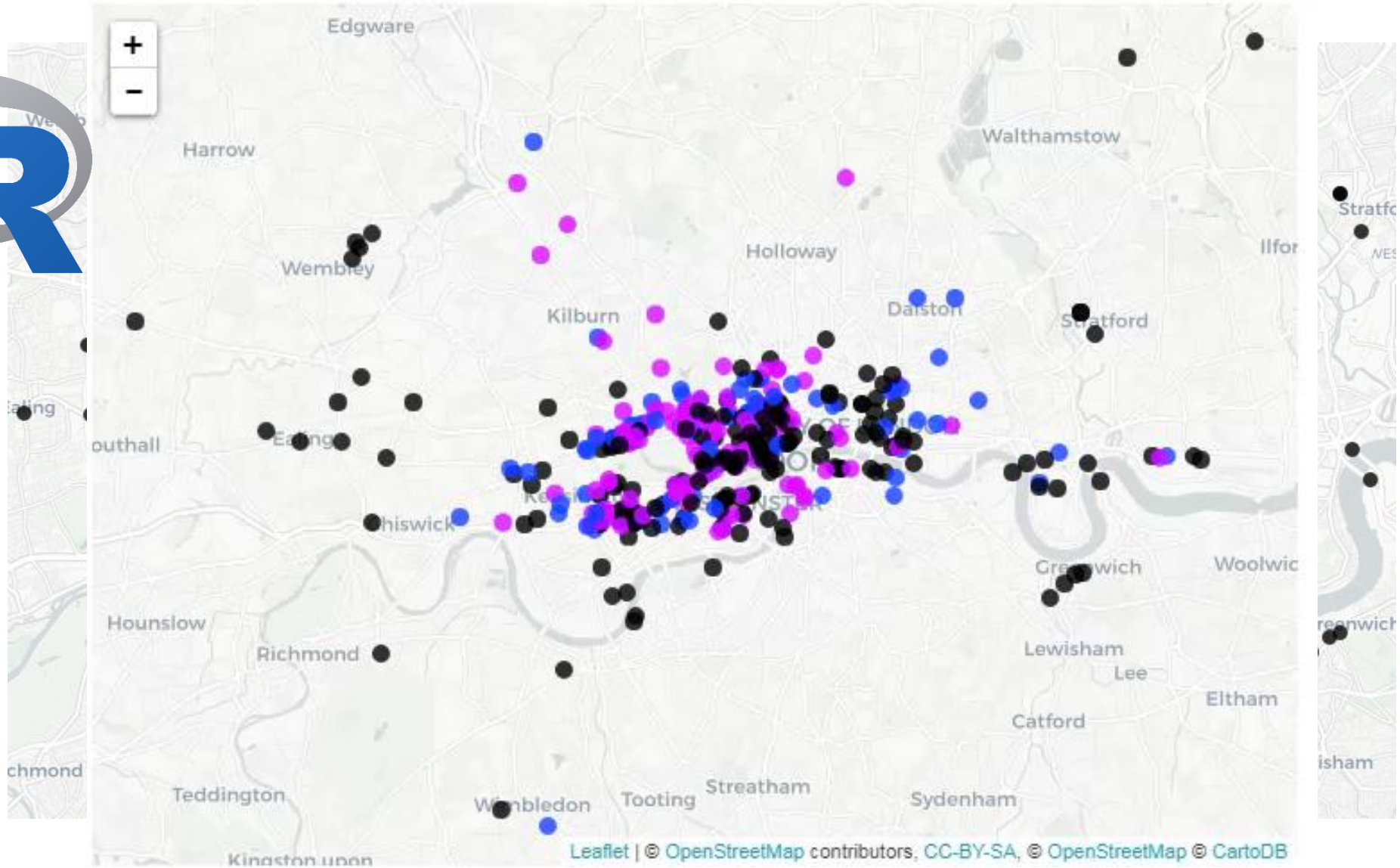


# CLUSTERING. **METHOD**



# CLUSTERING. **ON MAP**



# CLUSTERING. **BUSINESS OBJECTIVES MET**

Which business problems / predictions we could fulfill?

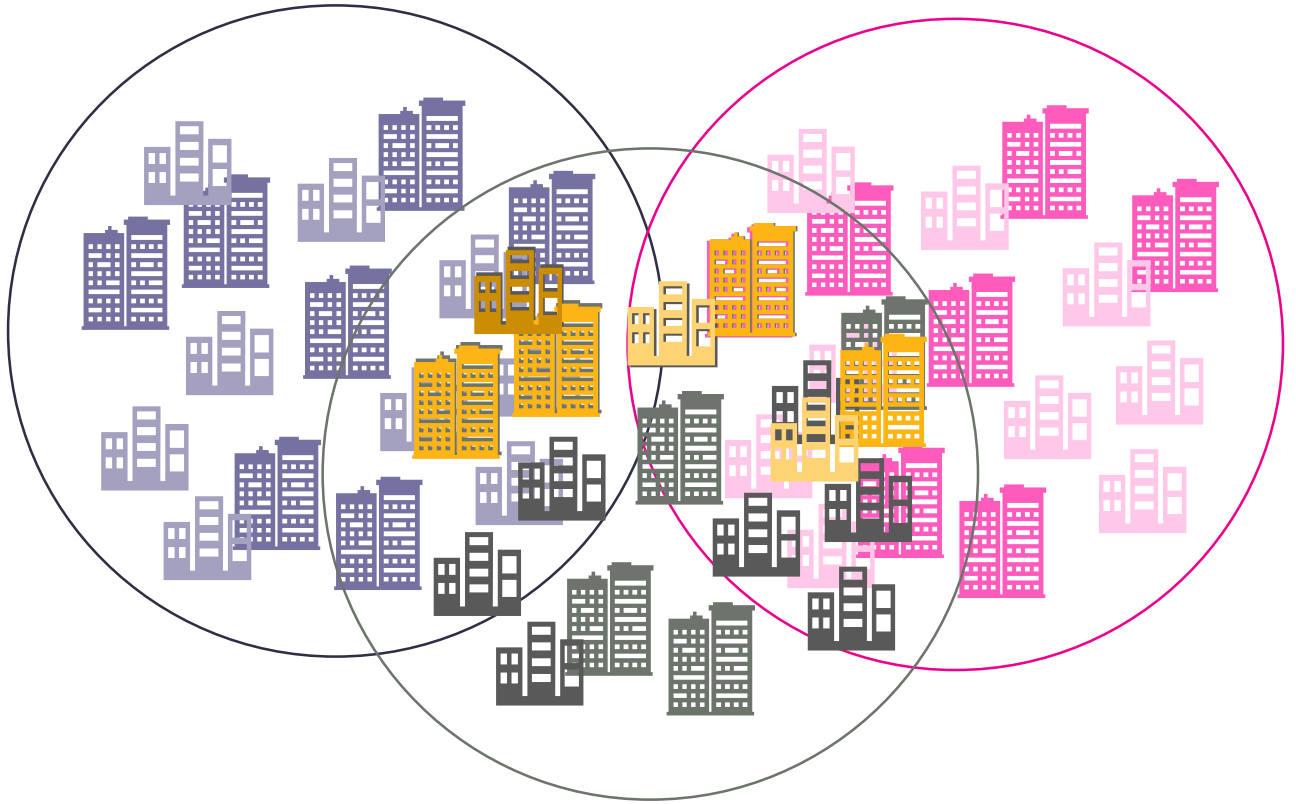
1. Machine Learning model (clustering) to **personalize search results**

2. **Prediction model for Segmentation** correlated to geo-location

3. **Recommendation engine for hotels by proximity**

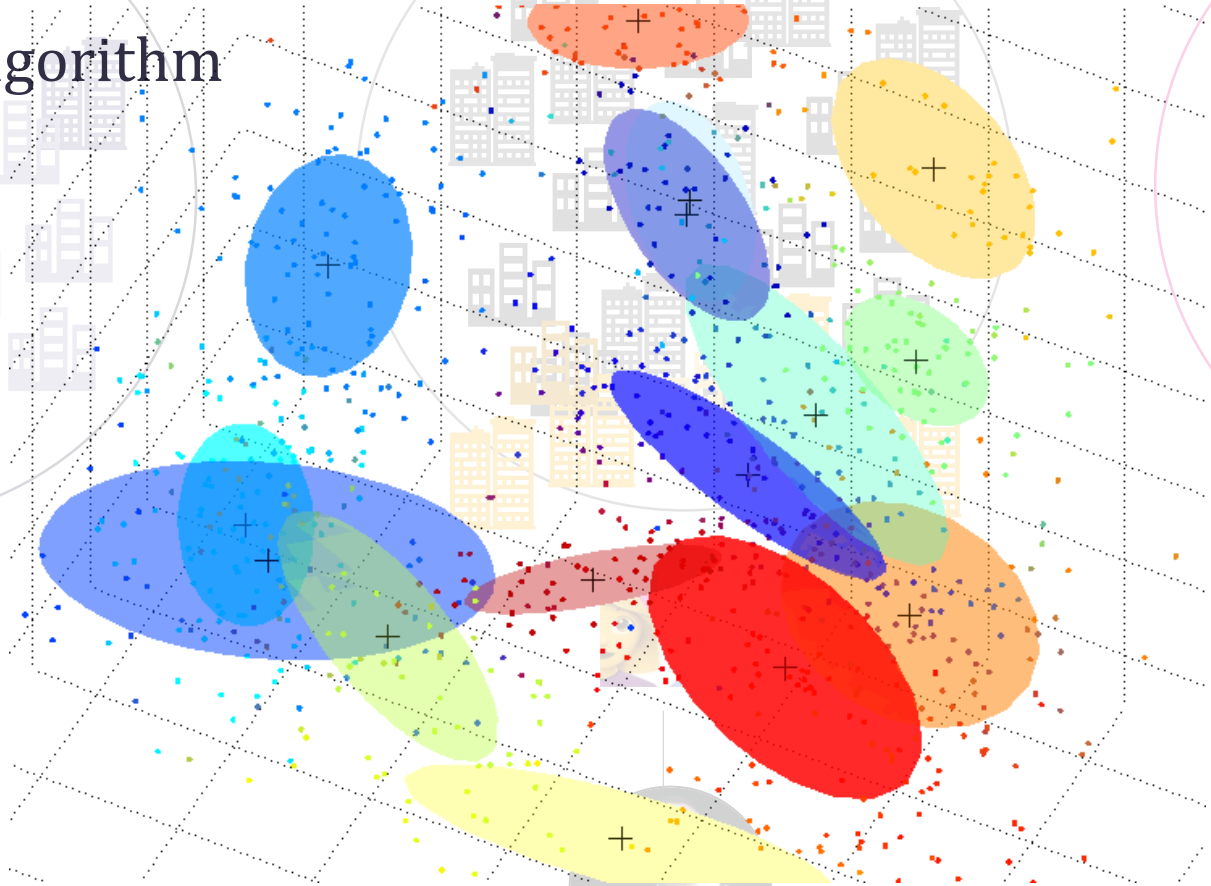
# CLUSTERING. **BUSINESS SOLUTION – REVIEW MODEL**

What are we missing with the previous algorithm?



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

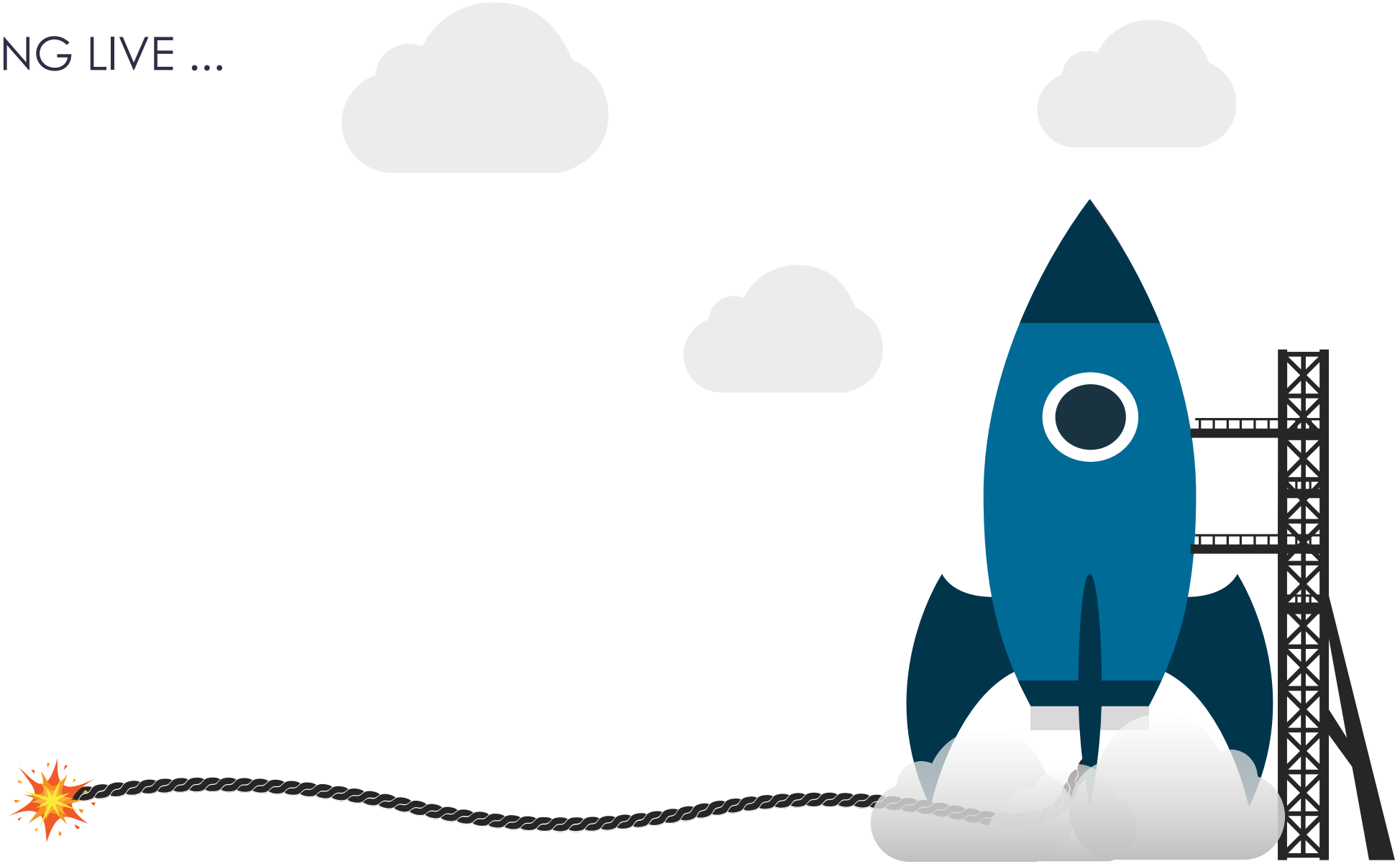
Maximization of the search results using **Gaussian Mixture Model** (GMM) through EM algorithm



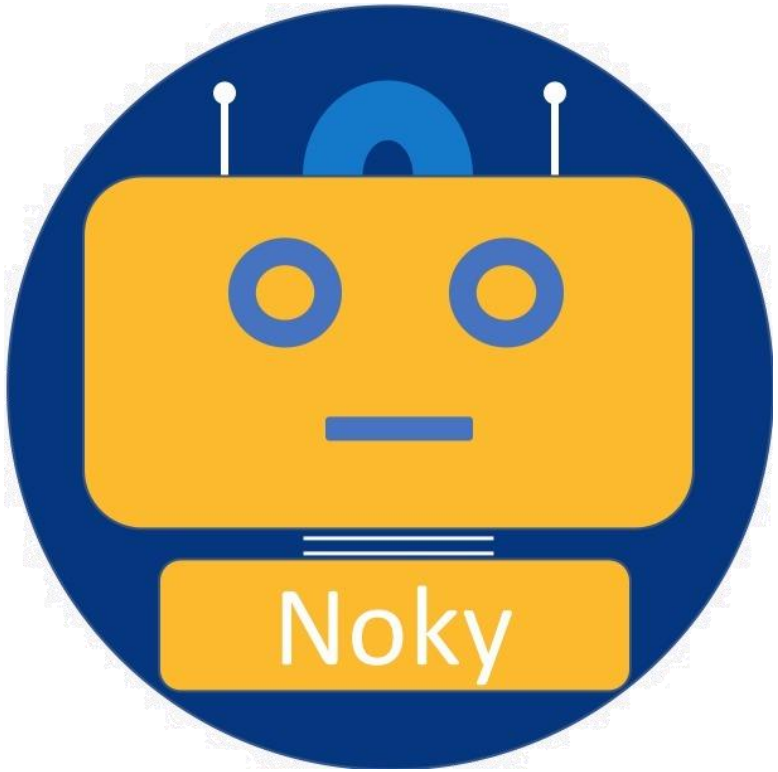
# NLP. BAG OF WORDS

# NLP. BAG OF WORDS

# GOING LIVE ...

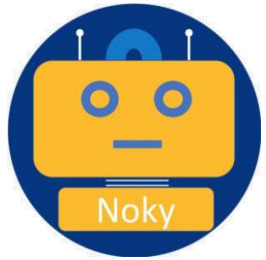


# INTO ACTION...




## 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

Integration to Open Source API

Clustered data to predict and optimize questions asked by bot

**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

**Recommended Hotels**

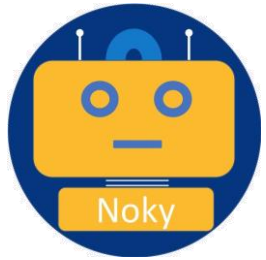
No

Feedback loop to optimize results

Yes

Continue for booking at **Booking.com**

# BUSINESS APPLICATION. **CHATBOT**




So, how can y

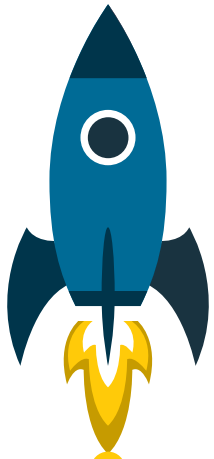
Wher

**Traveling** from Israel

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

ave **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

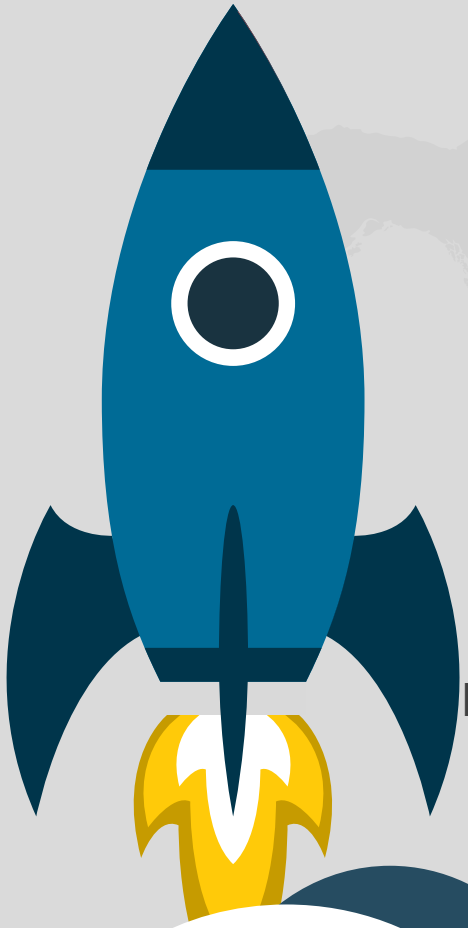
## DATA CORRELATIONS

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

IDC
MBA BIG DATA


ENJOY YOUR STAY