Solving Cold Start Problem For New Booking Aggregator

Dmitry Beresnev d.beresnev@innopolis.university

Vsevolod Klyushev v.klyushev@innopolis.university

Nikita Yaneev n.yaneev@innopolis.university

GitHub: https://github.com/Kadaverciant/bd-s25-project/

Introduction

Challenge

The online travel accommodation market is highly competitive, and launching a new booking aggregator in a city like Rio de Janeiro presents a critical hurdle: the cold start problem. Without historical user data—bookings, reviews, or behavioral insights—new platforms struggle to rank listings effectively or personalize recommendations, leading to poor user engagement and slow growth.

Our solution

We propose a data-driven approach leveraging publicly available Airbnb data to predict review scores for new, unrated listings. By analyzing features like location, price, amenities, and host attributes, our model generates reliable quality estimates, enabling:

1. Better ranking algorithms from day one
2. Enhanced discoverability of high-quality stays
3. Improved user experience and conversion rates

Introduction

Business Impact

This initiative directly addresses the cold start challenge, driving early-stage bookings, revenue growth, and market penetration. In this presentation, we'll explore our methodology, data insights, and the predictive system designed to ensure a strong launch.

airbnb

Data collection & Description

Dataset Overview:

- Source: Airbnb (Rio de Janeiro) via Kaggle (public dataset)
- Timeframe: 3 years (2018–2020)
- Size: ~784K records, 108 attributes (2.5 GB)

Key Features:

- Geodata: Housing coordinates
- Rental Attributes: Wi-Fi, TV, beds, etc.
- Host Characteristics: Host-related metrics
- Price: Rental price per listing
- Review Score Rating: Target variable for cold-start prediction

Data Preprocessing Pipeline

Feature Selection:

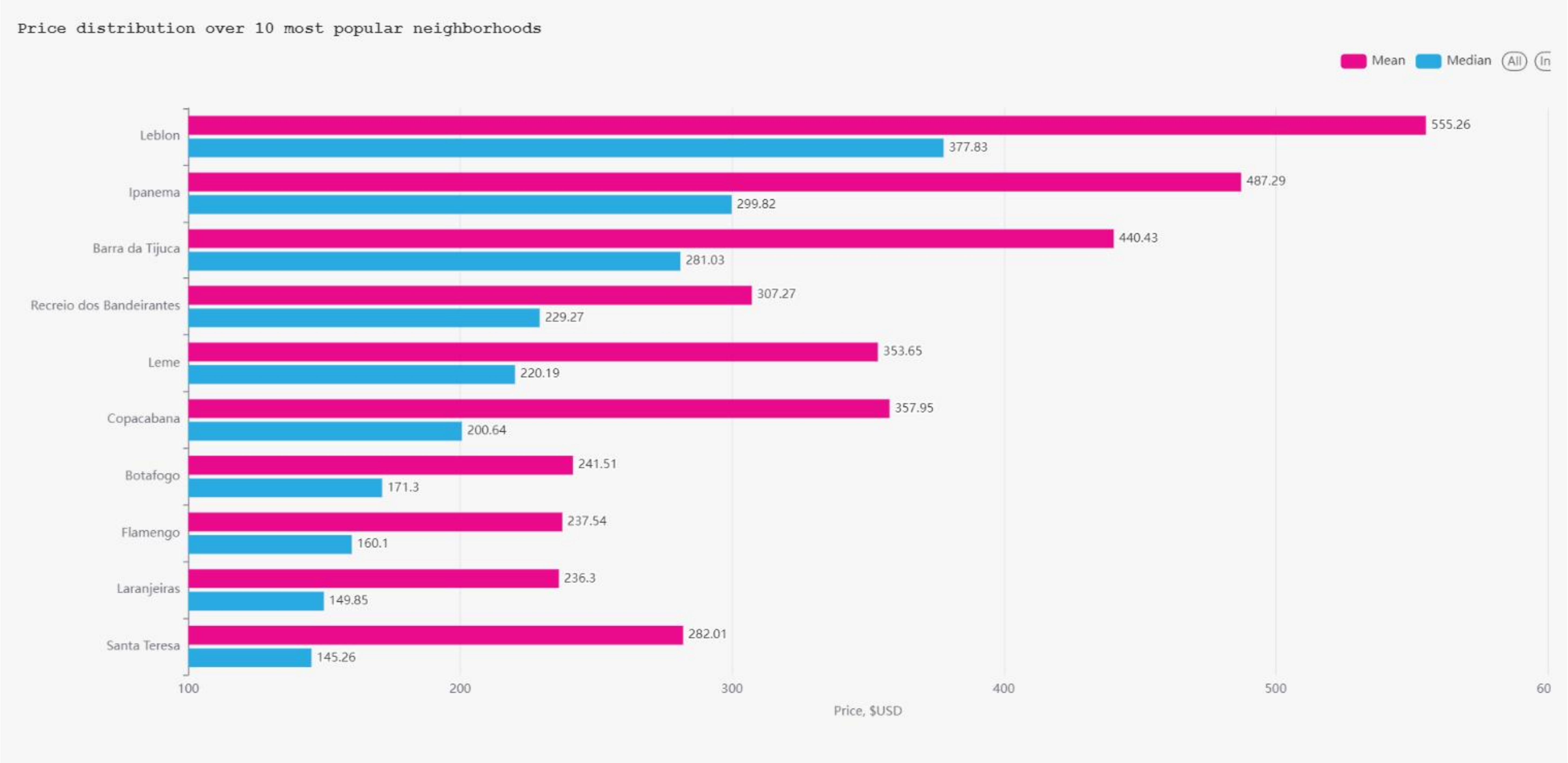
- Original: 108 columns → Selected: 29 key features
- Criteria: Relevance to rating prediction (e.g., host behavior, property details, pricing)
- Target: review_scores_rating

Data Cleaning:

- Handling Missing Values:
- Formatting
- Result: ~379,600 records retained

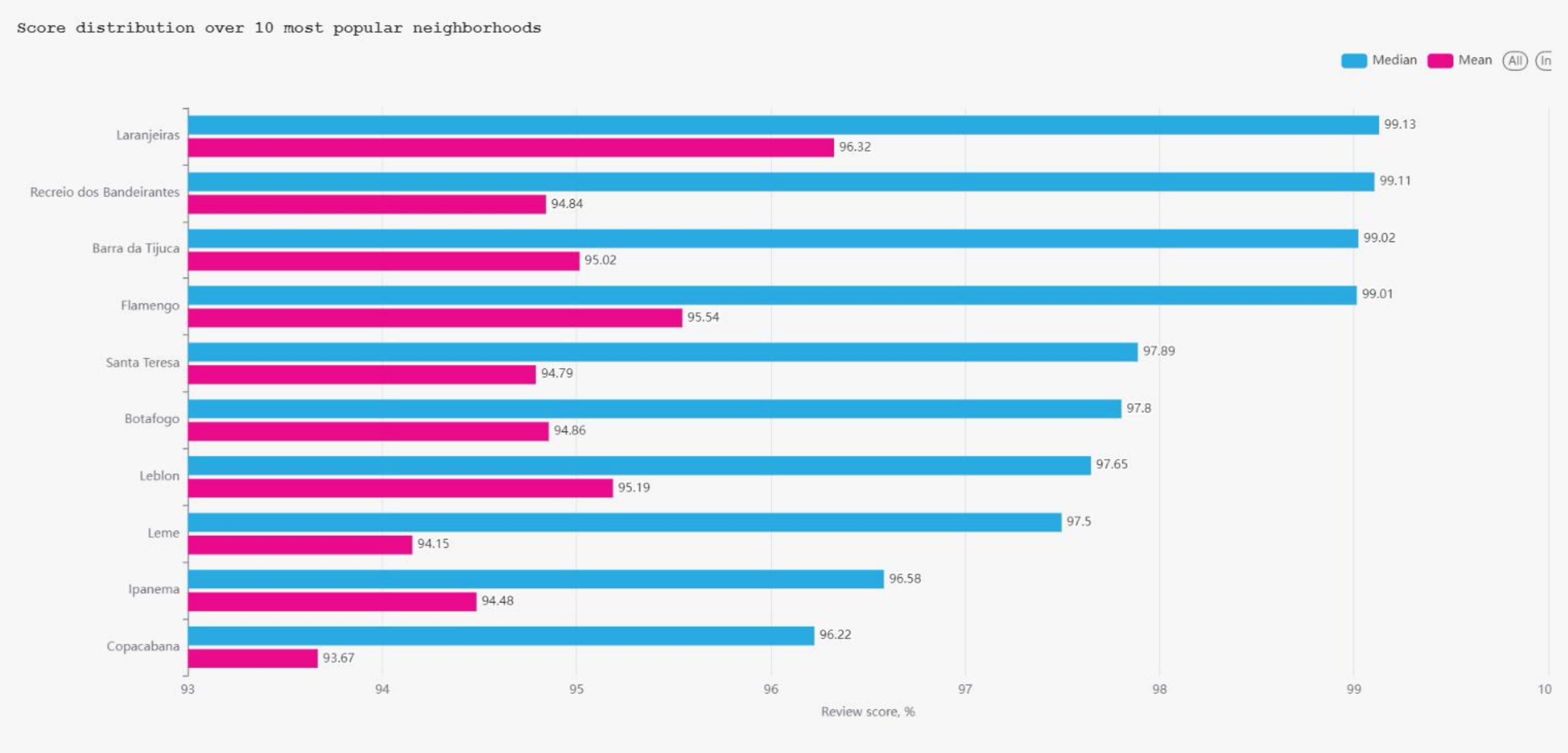
Feature Engineering:
The amenities field originally contained over 180 unique entries, representing various features of each listing. We selected the 20 most common ones in order to have more actionable information

Insight 1.1

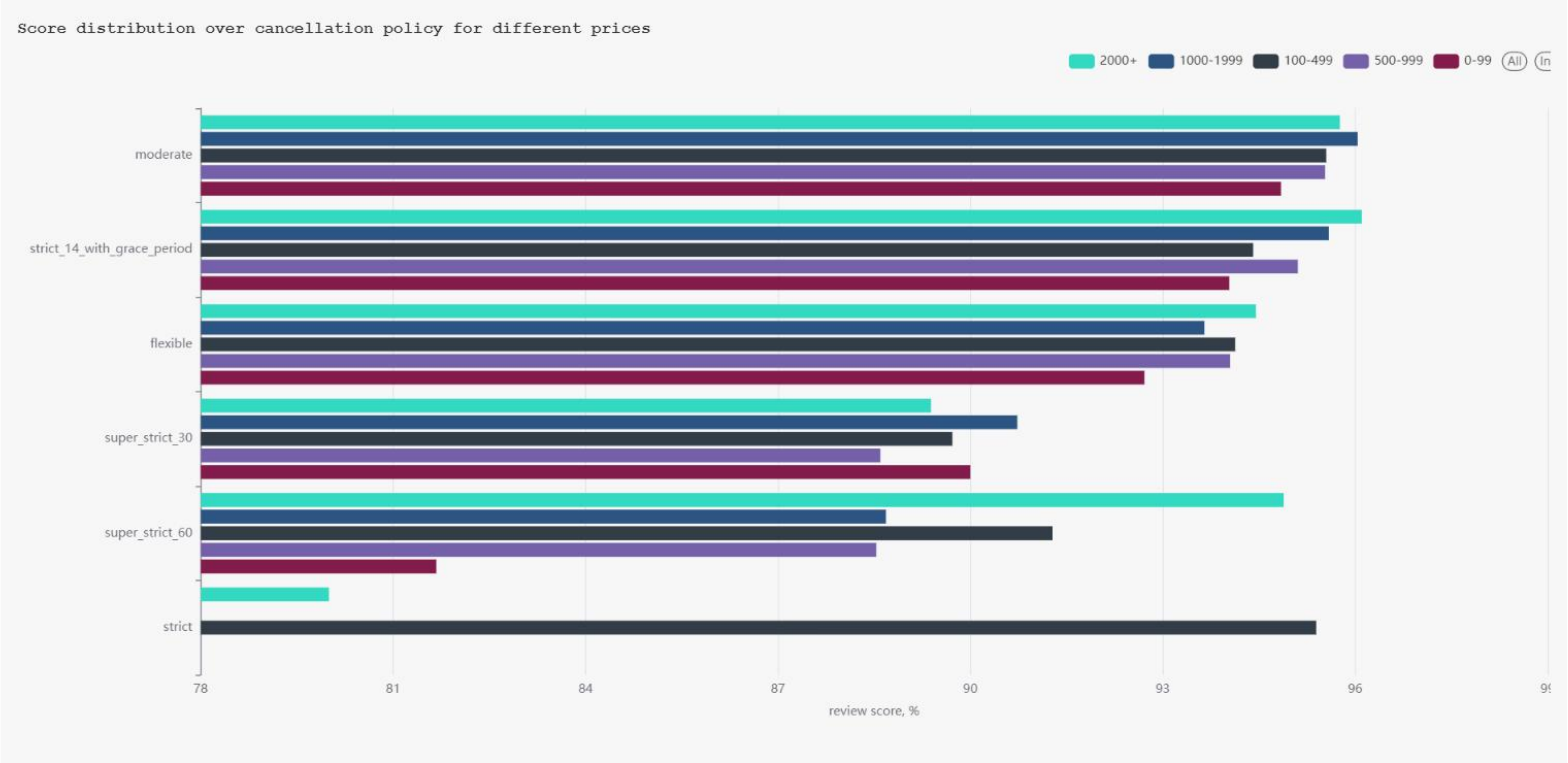


Price distribution over 10 most popular neighborhoods
Mean
Median
All
In
Leblon
555.26
377.83
Ipanema
487.29
299.82
Barra da Tijuca
440.43
281.03
Recreio dos Bandeirantes
307.27
229.27
Leme
353.65
220.19
Copacabana
357.95
200.64
Botafogo
241.51
171.3
Flamengo
237.54
160.1
Laranjeiras
236.3
149.85
Santa Teresa
282.01
145.26
100
200
300
400
500
60
Price, $USD

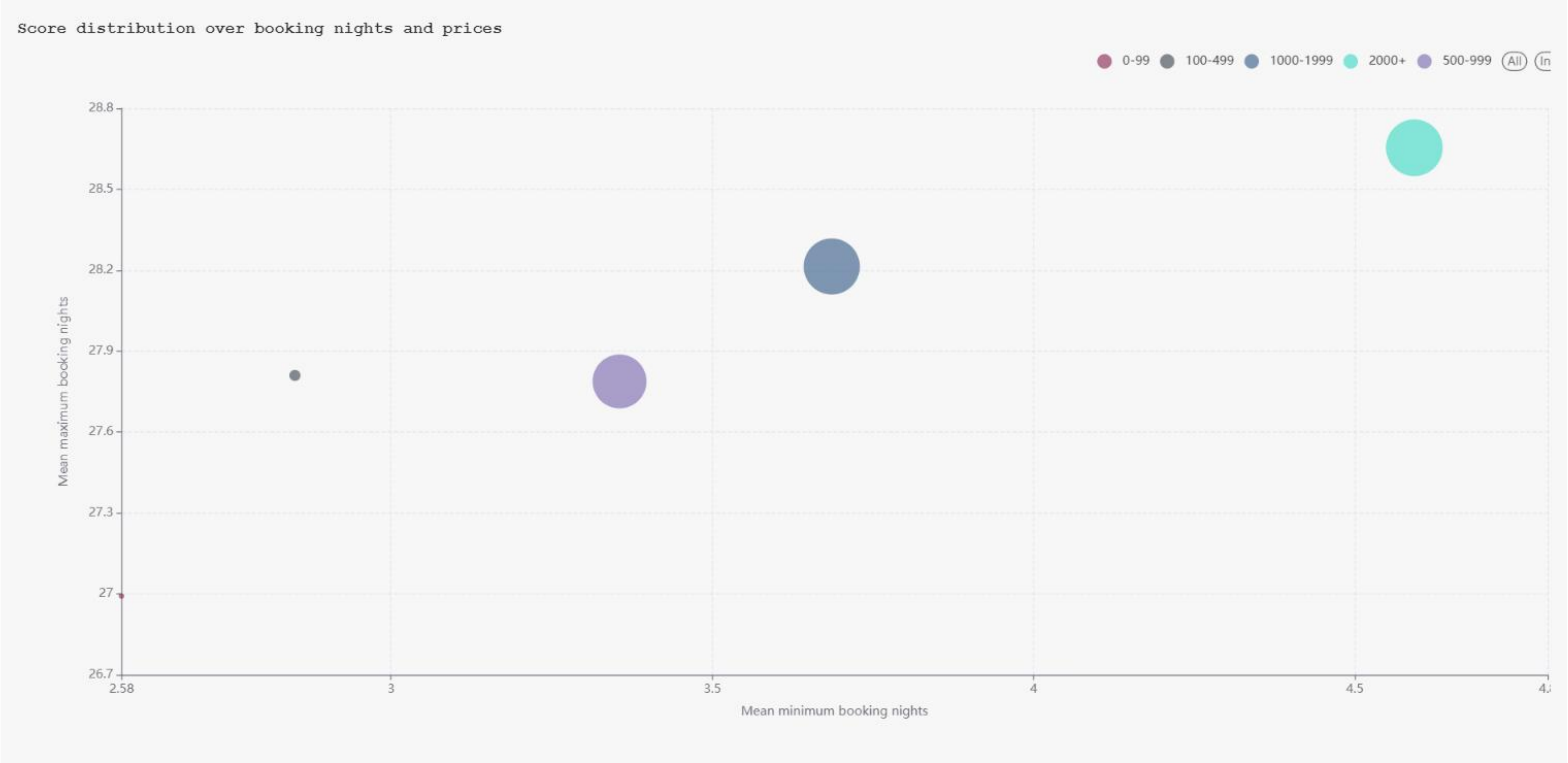
Insight 1.2


Score distribution over 10 most popular neighborhoods
Median
Mean
(All)
(In
Laranjeiras
Recreio dos Bandeirantes
Barra da Tijuca
Flamengo
Santa Teresa
Botafogo
Leblon
Leme
Ipanema
Copacabana
99.13
96.32
99.11
94.84
99.02
95.02
99.01
95.54
97.89
94.79
97.8
94.86
97.65
95.19
97.5
94.15
96.58
94.48
96.22
93.67
93
94
95
96
97
98
99
10
Review score, %

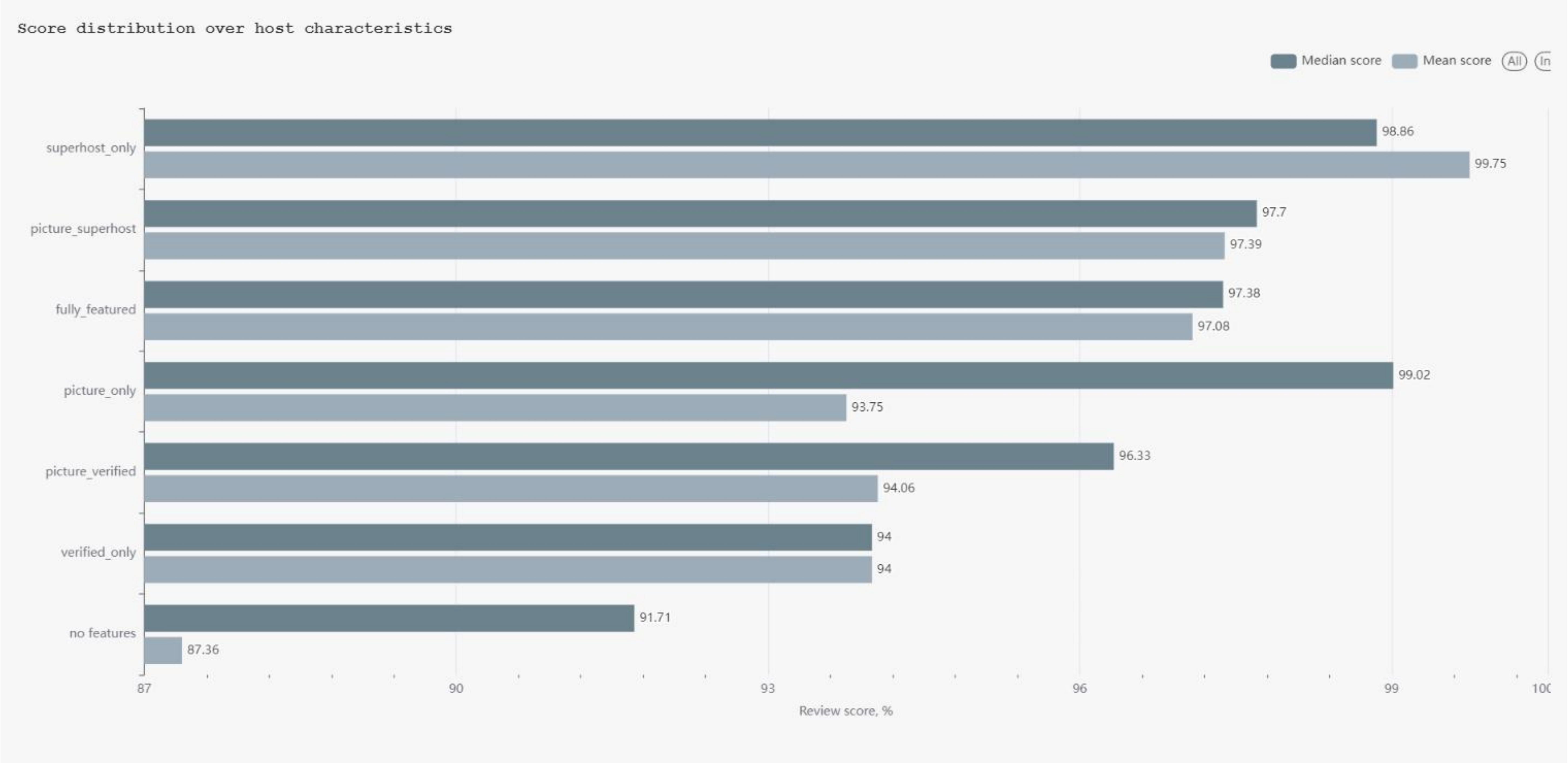
Insight 2.1


Score distribution over cancellation policy for different prices
2000+
1000-1999
100-499
500-999
0-99
All
moderate
strict_14_with_grace_period
flexible
super_strict_30
super_strict_60
strict
78
81
84
87
90
93
96
review score, %

Insight 2.2


Score distribution over booking nights and prices
0-99
100-499
1000-1999
2000+
500-999
All
28.8
28.5
28.2
27.9
27.6
27.3
27
26.7
Mean maximum booking nights
2.58
3
3.5
4
4.5
Mean minimum booking nights

Insight 3

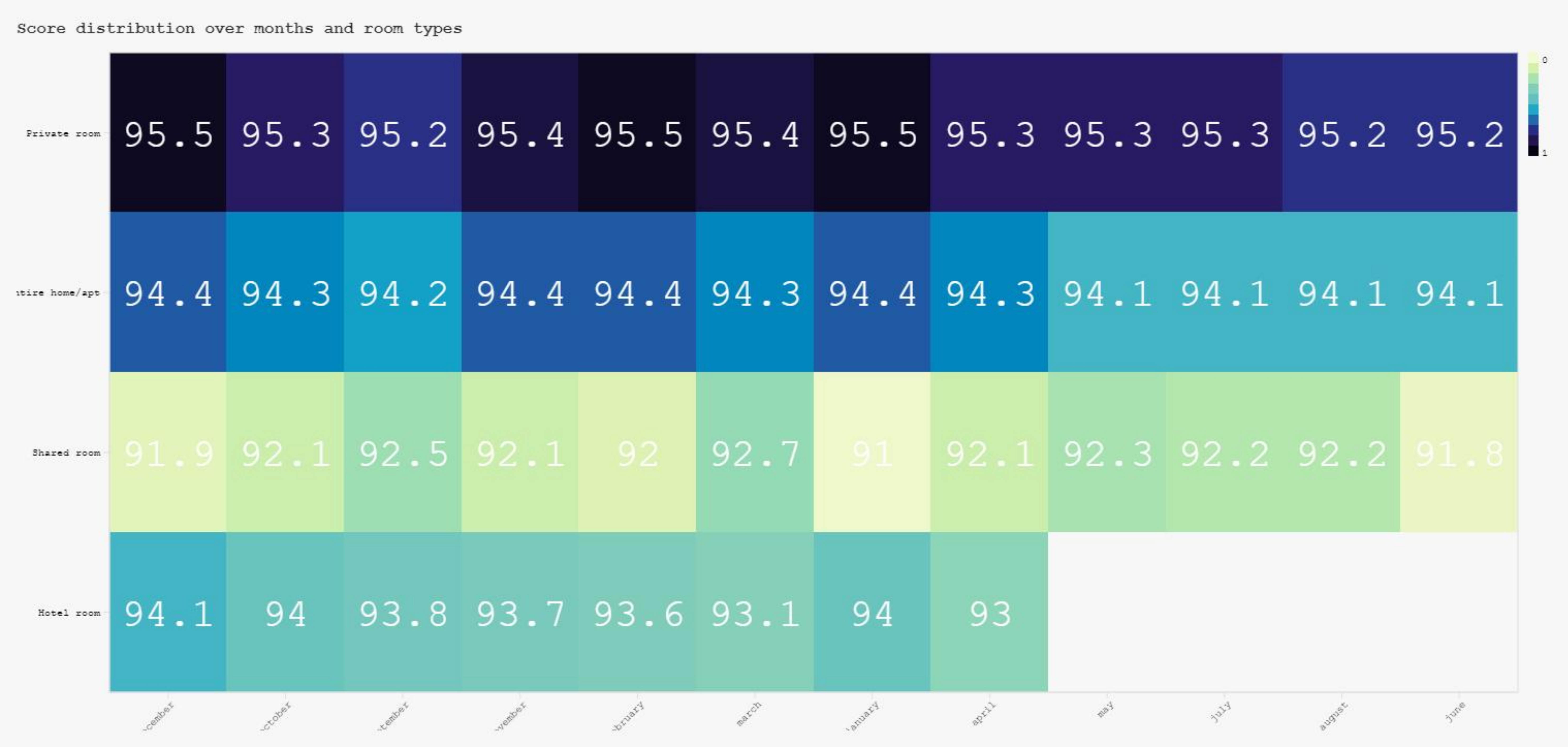


Score distribution over host characteristics
Median score
Mean score
All
superhost_only
98.86
99.75
picture_superhost
97.7
97.39
fully_featured
97.38
97.08
picture_only
99.02
93.75
picture_verified
96.33
94.06
verified_only
94
94
no features
91.71
87.36
87
90
93
96
99
Review score, %

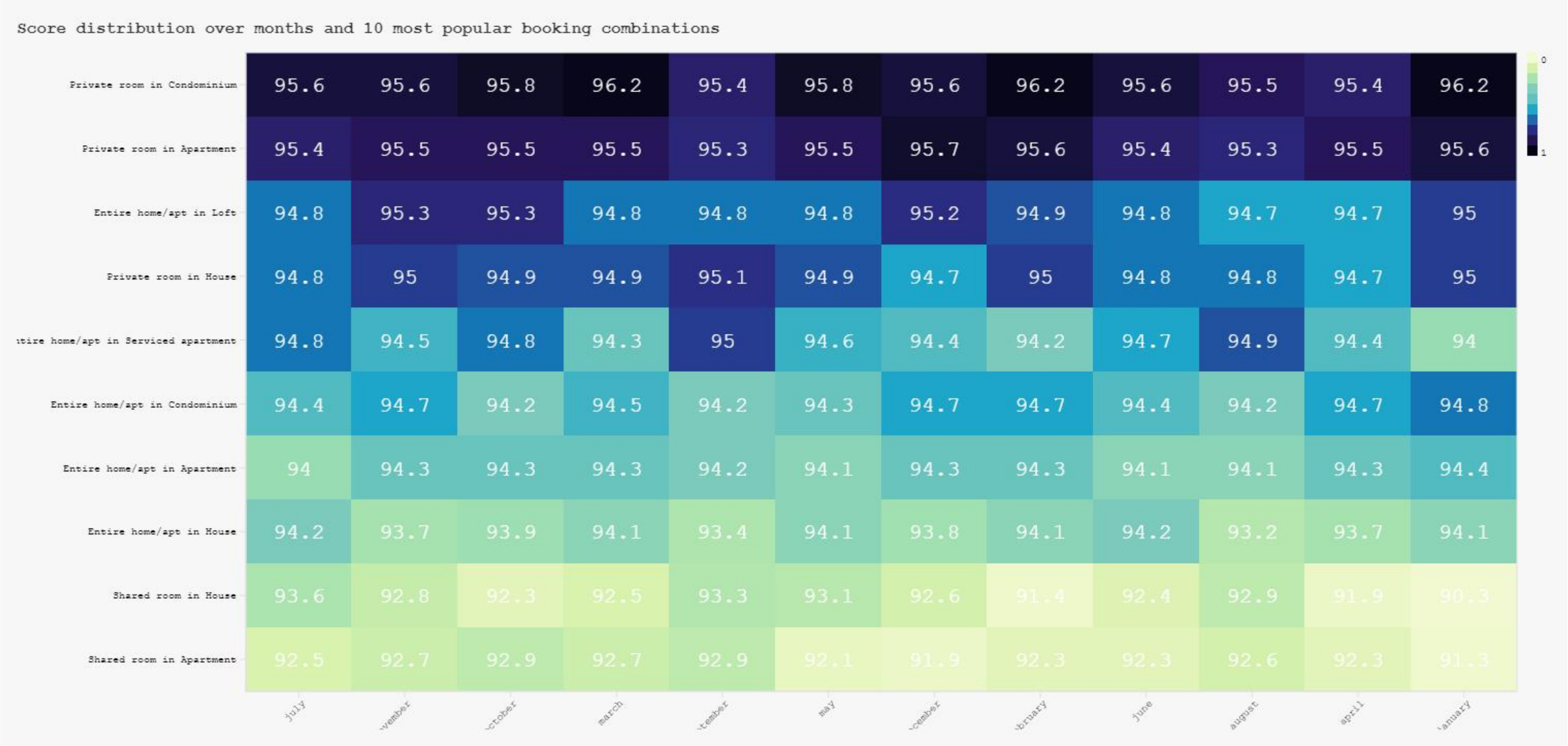
Insight 4.1


Score distribution over months and 10 most popular property types
Guest suite
Loft
Townhouse
Guesthouse
Condominium
Apartment
erviced apartment
House
Hostel
ed and breakfast
bruary
cember
may
march
anuary
vember
july
tember
april
august
june
ctober
0
1
95.4 95.7 95.1 95 95.3 95.5 95.1 95.3 95.1 95 95.1 95.3
95 95.3 94.9 94.8 95.1 95.4 95 94.9 94.8 94.7 95 95.3
95.3 94.9 94.7 95.4 95.4 95 94.9 94.4 95.7 94.6 95.1 94.7
95.2 95 94.8 95.2 94.5 94.2 95.2 95 95.4 95.1 95 94.8
95 94.9 94.7 94.9 95.1 94.9 94.7 94.5 94.9 94.5 94.6 94.6
94.6 94.6 94.4 94.5 94.6 94.6 94.3 94.4 94.5 94.4 94.4 94.6
94.3 94.3 94.4 94.3 94 94.4 94.6 94.8 94.5 94.7 94.6 94.7
94.4 94.2 94.5 94.4 94.3 94.4 94.5 94.3 94.1 94.1 94.4 94.3
91.1 91.6 93.9 91.6 90.9 92.9 93.5 93.5 91.2 93.4 93.6 93.2
94.1 93.2 92 93.1 93.7 91.6 91.1 91.4 92 91.6 90.1 90.2

Insight 4.2


Score distribution over months and room types
Private room
95.5 95.3 95.2 95.4 95.5 95.4 95.5 95.3 95.3 95.3 95.2 95.2
ntire home/apt
94.4 94.3 94.2 94.4 94.4 94.3 94.4 94.3 94.1 94.1 94.1 94.1
Shared room
91.9 92.1 92.5 92.1 92 92.7 91 92.1 92.3 92.2 92.2 91.8
Hotel room
94.1 94 93.8 93.7 93.6 93.1 94 93
cember
ctober
tember
vember
bruary
march
anuary
april
may
july
august
june
0
1

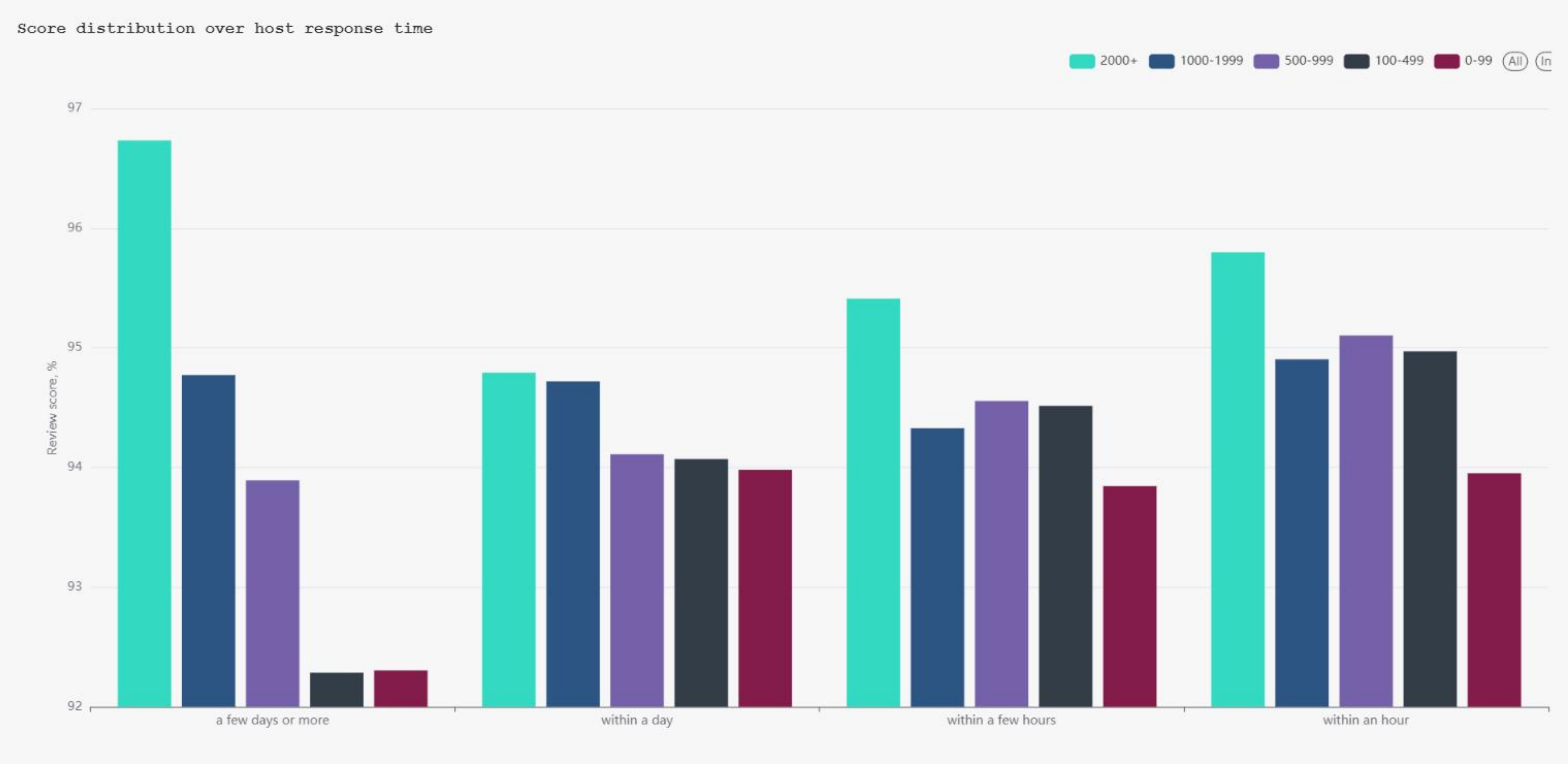
Insight 4.3


Score distribution over months and 10 most popular booking combinations
Private room in Condominium
95.6 95.6 95.8 96.2 95.4 95.8 95.6 96.2 95.6 95.5 95.4 96.2
Private room in Apartment
95.4 95.5 95.5 95.5 95.3 95.5 95.7 95.6 95.4 95.3 95.5 95.6
Entire home/apt in Loft
94.8 95.3 95.3 94.8 94.8 94.8 95.2 94.9 94.8 94.7 94.7 95
Private room in House
94.8 95 94.9 94.9 95.1 94.9 94.7 95 94.8 94.8 94.7 95
ntire home/apt in Serviced apartment
94.8 94.5 94.8 94.3 95 94.6 94.4 94.2 94.7 94.9 94.4 94
Entire home/apt in Condominium
94.4 94.7 94.2 94.5 94.2 94.3 94.7 94.7 94.4 94.2 94.7 94.8
Entire home/apt in Apartment
94 94.3 94.3 94.3 94.2 94.1 94.3 94.3 94.1 94.1 94.3 94.4
Entire home/apt in House
94.2 93.7 93.9 94.1 93.4 94.1 93.8 94.1 94.2 93.2 93.7 94.1
Shared room in House
93.6 92.8 92.3 92.5 93.3 93.1 92.6 91.4 92.4 92.9 91.9 90.3
Shared room in Apartment
92.5 92.7 92.9 92.7 92.9 92.1 91.9 92.3 92.3 92.6 92.3 91.3
july
vember
ctober
march
tember
may
cember
bruary
june
august
april
anuary
0
1

Insight 4.4

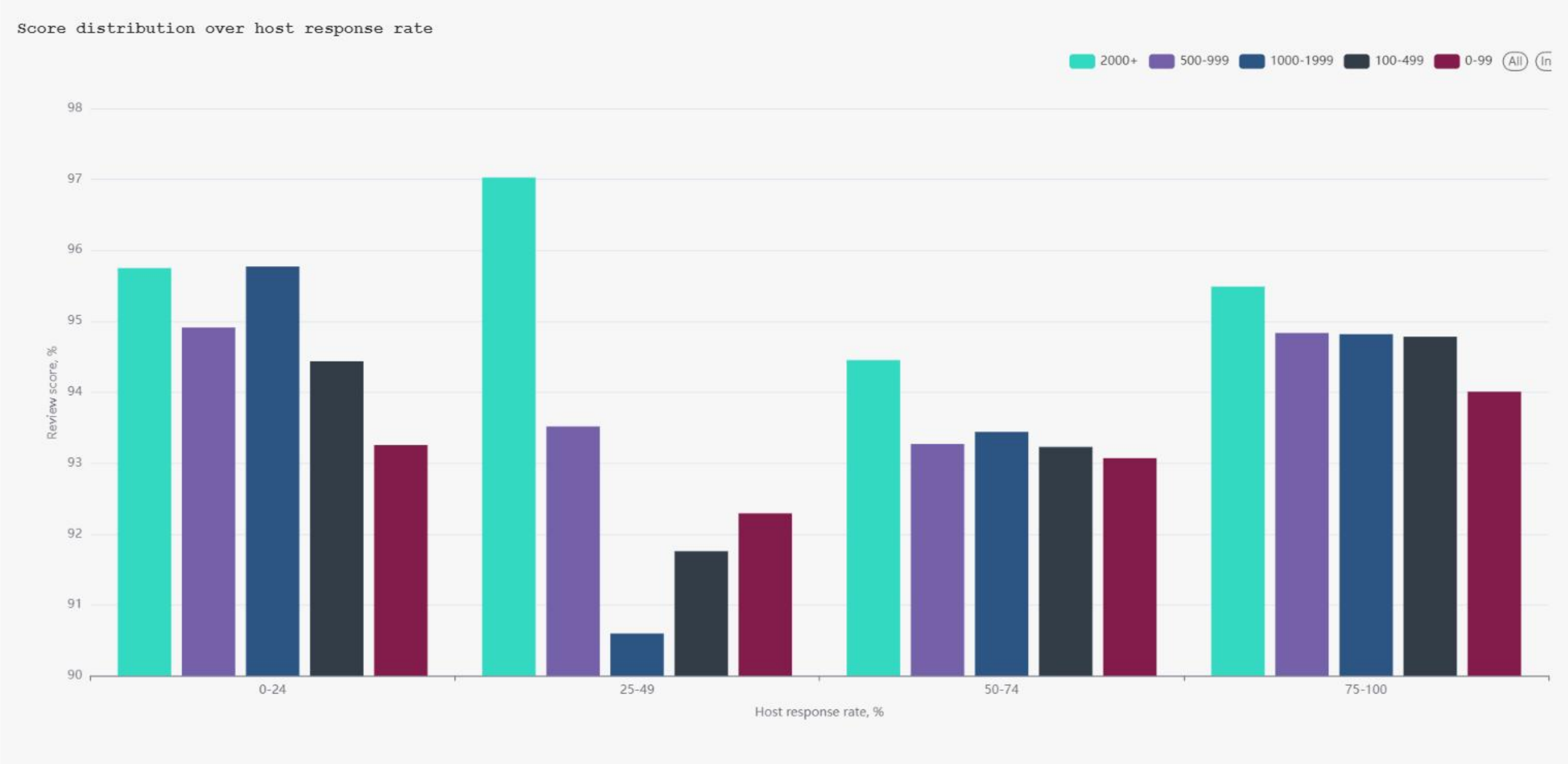

Price distribution over months and 10 most popular booking combinations
Entire home/apt in House
959 897 958 910 964 905 883 872 880 871 866 859
Entire home/apt in Apartment
464 436 421 459 410 415 413 410 420 417 408 410
Entire home/apt in Condominium
454 419 397 398 390 405 383 379 394 400 388 391
itire home/apt in Serviced apartment
386 390 381 363 379 379 364 364 341 357 358 350
Entire home/apt in Loft
334 325 365 330 377 345 387 379 331 311 364 328
Private room in Apartment
188 189 183 187 181 185 181 181 184 186 179 181
Private room in Condominium
183 178 158 175 152 157 156 169 173 173 156 171
Private room in House
163 161 159 162 156 157 154 155 162 161 153 156
Shared room in Apartment
145 139 116 136 116 120 127 133 141 141 123 140
Shared room in House
97 94 87 97 89 90 88 93 95 95 87 95
april
march
june
bruary
july
may
tember
ctober
cember
anuary
august
vember
0
1

Insight 5.1



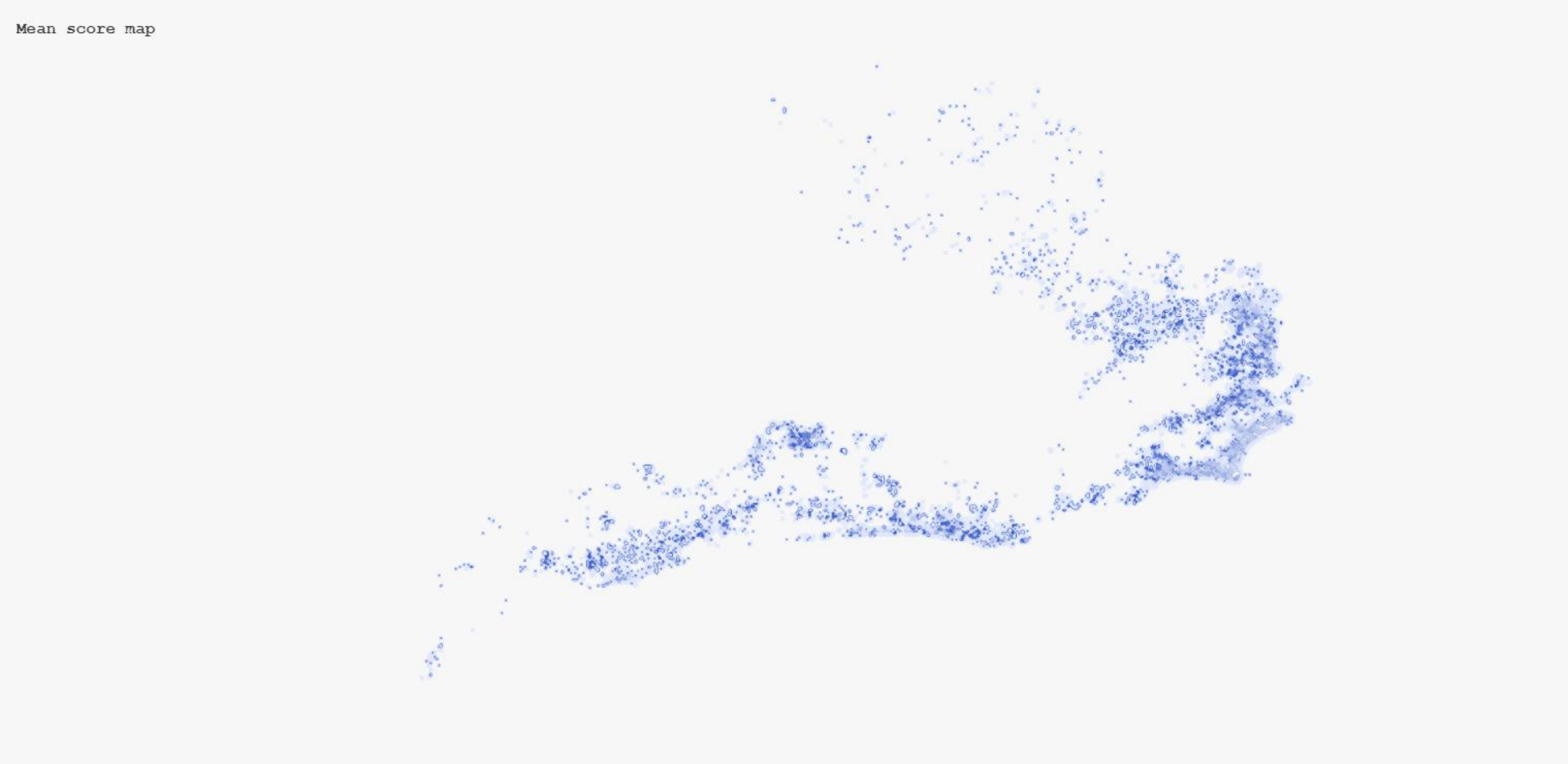
Score distribution over host response time
2000+
1000-1999
500-999
100-499
0-99
All
97
96
95
94
93
92
Review score, %
a few days or more
within a day
within a few hours
within an hour

Insight 5.2

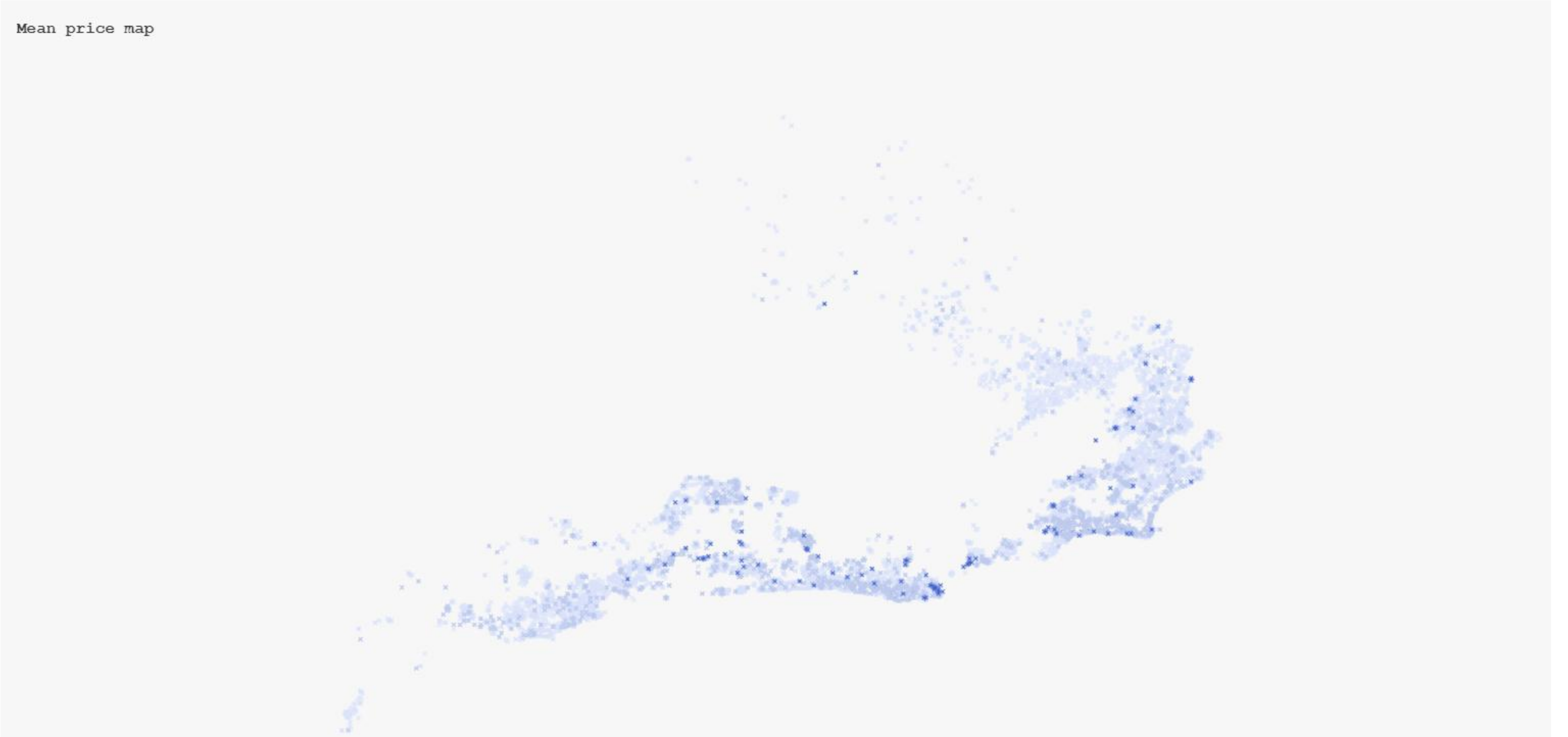


Score distribution over host response rate
2000+
500-999
1000-1999
100-499
0-99
All
98
97
96
95
94
93
92
91
90
Review score, %
0-24
25-49
50-74
75-100
Host response rate, %

Insight 6.1

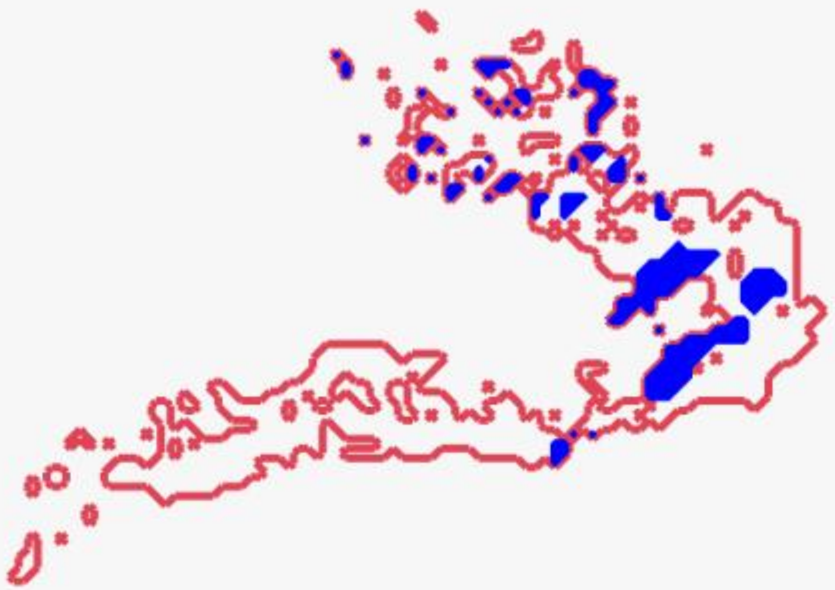


Insight 6.2



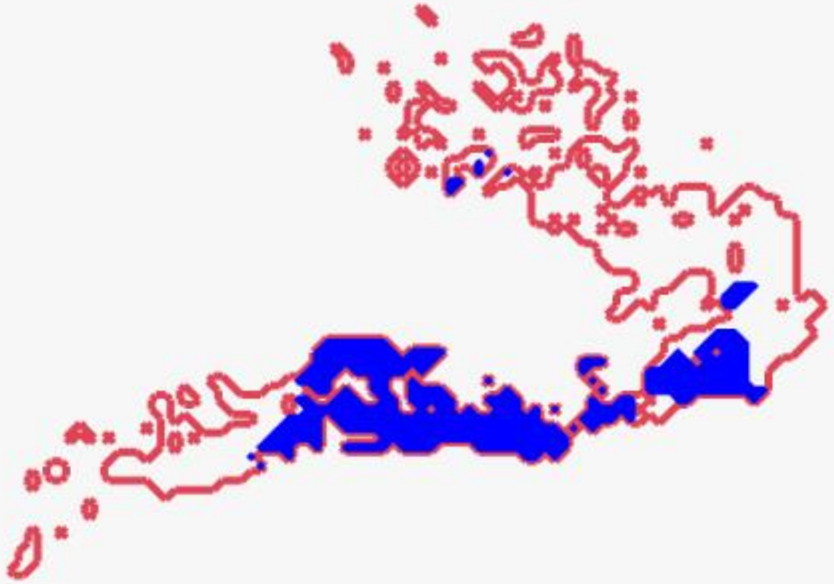
Insight 6.3

Top neighborhoods by mean score



Insight 6.4

Top neighborhoods by mean price



Insight 7

Score distribution over amenities and months
wifi_ac_fridge 96 95.8 95.8 95.9 95.9 95.1 96 95.9 95.1 95.6 95.5 95.1
all_amenities 95.4 95.4 95.2 95.3 95.5 95.1 95.4 95.4 95.1 95.3 95.3 95.1
wifi_ac 94.7 94.9 94.6 94.6 94.8 94.9 94.6 94.7 94.9 94.9 94.8 94.8
wifi_ac_hotwater 94.6 94.6 94.5 94.6 94.6 94.6 94.6 94.6 94.6 94.5 94.5 94.6
wifi_fridge 94.8 95 94.3 94.3 95.1 93.6 94.9 95 93.8 94.2 94 94
ifi_hotwater_fridge 94 94 94 94.3 93.8 94 94.1 93.9 93.7 93.8 93.8 94
wifi_hotwater 93.4 93.2 93.2 93.4 93.2 93.1 93.5 93.5 93 92.9 92.8 93.1
wifi_only 92.9 93.1 93 93.1 93 92.7 93 92.9 92.7 93 93.1 92.9
ac_fridge 93.4 91.6 93 92.9 93.1 92.3 93.1 92.4 92.2 92.1 92.2 91.4
ac_hotwater 92.2 92.3 92.2 92.5 92.4 92.3 92.5 92.2 92.3 92.2 92.3 92.6
ac_hotwater_fridge 92.2 91.9 92.1 92 92.2 91.9 91.6 92.1 92.4 91.8 91.8 92.2
ac_only 91.5 92 91.8 91.3 91.8 91.4 91.2 91.6 92.1 91.3 91.1 92.7
fridge_only 92.9 91.2 91.9 91.6 91.8 90.4 91.2 92.4 89 91.6 91.2 88.7
no_amenities 91.4 91.5 90.6 91.2 91.7 90.5 91.2 92 90.6 90.9 91.1 91.2
hotwater_fridge 89.3 92.1 91.1 89.8 88.6 93.1 88.7 88.8 92.7 91.9 91.6 89.7
hotwater_only 90.7 89.9 91.1 91.1 90.3 91.2 90.4 90.5 91.1 88.7 88.6 91.4
january
ctober
april
march
vember
july
bruary
cember
june
tember
august
may
0
1

Model Selection & Performance

Why These Models?:

Linear Regression:

- Pros: Fast, interpretable, sets a performance baseline.
- Use Case: Understand key drivers of ratings (e.g., "Price impacts scores linearly").
- Params Grid Search: Reg. Params = [0.01, 0.1, 1] and Elastic Net = [0, 0.5, 1]

Random Forest:

- Pros: Captures complex patterns (e.g., "Superhost status + location boost ratings").
- Outperformed Linear Regression
- Params Grid Search: Num Trees = [5, 10] and Max Depth = [5, 10]

Model	Num Trees	Max Depth	Reg. Params	Elastic Net
Linear Regression	-	-	0.01	0.5
Random Forest	10	10	-	-

Model	RMSE	MAE
Linear Regression	9.06	5.68
Random Forest	8.28	5.34

Discussion

We successfully implemented predictive models, and they demonstrate strong performance

Key Results:

- Random Forest outperformed Linear Regression, achieving a lower MAE and reliably estimating listing ratings—even without prior review data.
- Linear Regression struggled with extreme/high scores, revealing limitations in handling complex patterns.

How Our Model Drives Business Success

- Higher Revenue & Customer Satisfaction
- Accurate Rating Predictions → Ensures new listings (with no reviews) get fair visibility, boosting bookings and trust.
- Reduces "Cold Start" Penalty → Helps high-quality listings rank faster, increasing platform revenue.

Challenges we are faced

We had not reformulated the problem itself during the work. Initially we decided to use only subset of the original dataset.

However, we faced with problem with big number of undefined value: after filtering them the dataset becomes too small to satisfy project requirements. So we have decided to use the whole initial dataset, so after filtering undefined values, it size remains acceptable.

Team Roles

Nikita Yaneev	Vsevolod Klyushev	Dmitry Beresnev
ML - engineer Prepared the data for the model, put together a full-fledged pipeline, trained the models, evaluated them, and identified the best solutions for our task.	Data engineer Implemented data collection, ingestion, preparation and storage pipeline. (Also wrote queries for data bucketing and partitioning.) Tested several compressing techniques and choose the best one.	Data analytic Performed EDA: built seven queries and analyzed the results; Created dashboards for Data Analysis, Insights and Solution models sections; Consult ML and Data engineers on approaches and solution choices based on data nature

Thanks for your attention