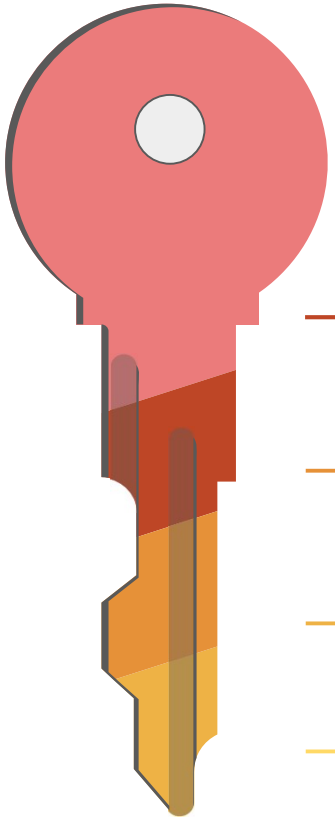
Booking behaviour

by argv1

» neue fische

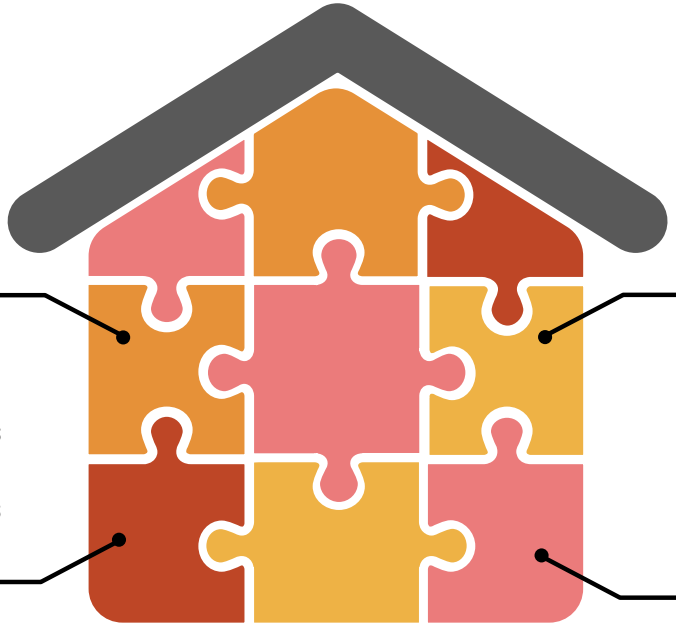
Agenda



1. **Personal motivation** — What drives us
2. **Our Goals** — Targets we would like to achieve
3. **DS lifecycle** — Used data, our approach and methods used
4. **Findings** — Visualising our insights
5. **Outlook** — What could comes next?

My personal motivation

Grown up
in the parental
hotel business



Logistics major
during my business
administration master

More than ten years
of experience in the
hospitality business

Still strongly interested
in the developments
within the industry.

Our Goals

VIPs

Research behavioural pattern of prospective customers (spend more than 75 quantile)

Repeater

Identify regular guest

Booking window

Forecast bookings on a quarterly basis

The Hotels & Data

Five houses in different locations — **The Hotels**

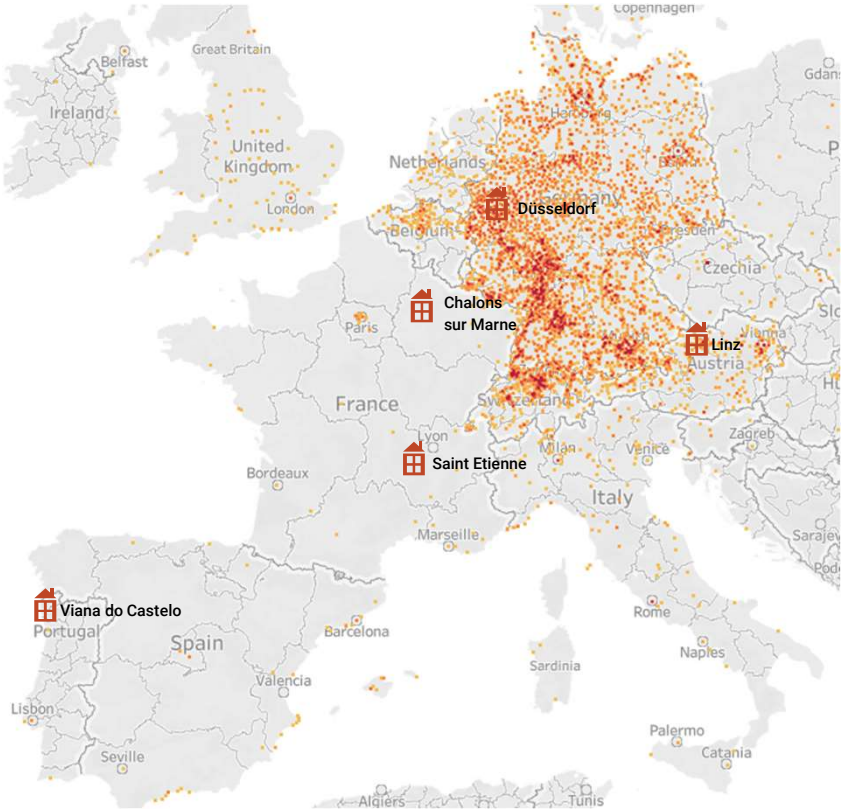
Booking, client and regional data sets — **Data Set**

115.119 individual reservations — **Reservations**

245.132 lines of bookings & client information — **The Clients**

8.209 lines of regional data — **Region**

Geographic Overview



Five hotel destinations

Customers plotted by density

Cleaning

START
245.132 rows

MISSINGS

1

IN THE VARIABLES
26% age category
19% distance
10% title, zip code, country
4% destination
area information outside Germany

2

IN GENERAL
Room capacity, CXLs, denails, events at the location description e.g. for additional services

WRONG VALUES

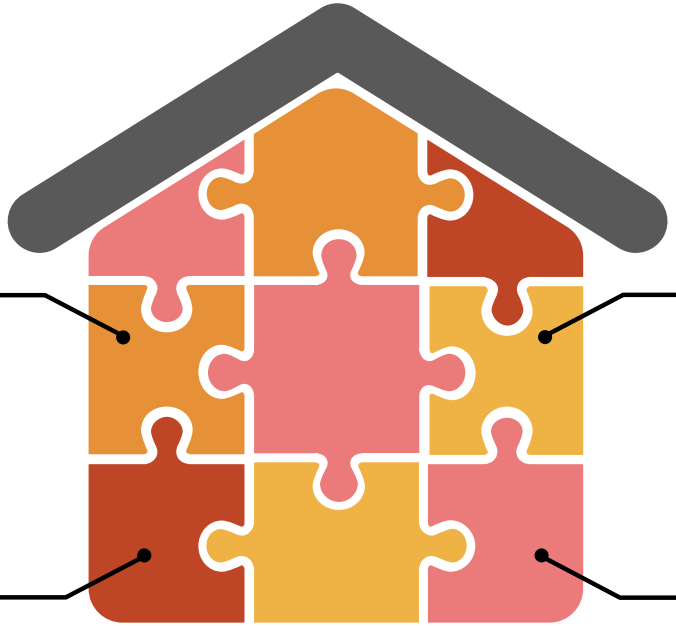
3

MISLEADING
anonymized destination, booking channel, etc.

4

INCONSISTENT
lines per booking vs. guest per booking (34%)
nights booked vs. departure - arrival date
arrival before booking

Feature engineering



Splitting
mostly timedate variables

Joining
booking channel, advertising ban by channel

Consistency
nights stay, booking rows

Additional
First booking, repeating guest, lead time, price per night, etc.

» neue fische

Demographics

Gender

52%
Women

48%
Men

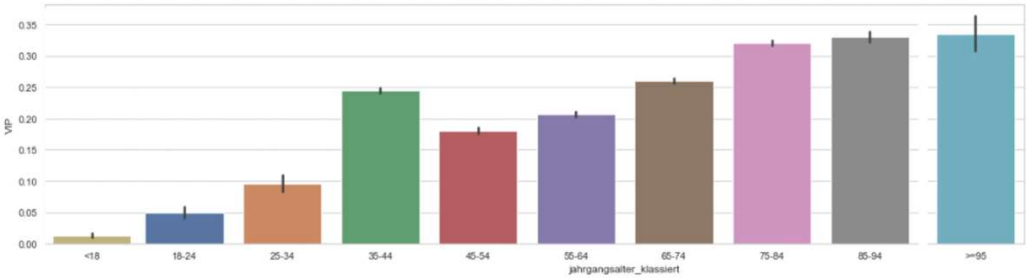
23%
65-74

22%
75-84

Nationalities

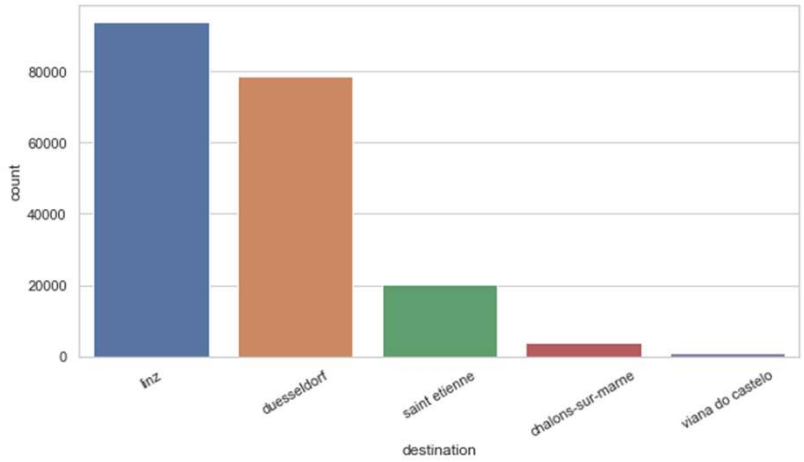
- 101 countries
- 94% from Germany
- 98,5% speaks German
- 13.125 cities

Age Groups

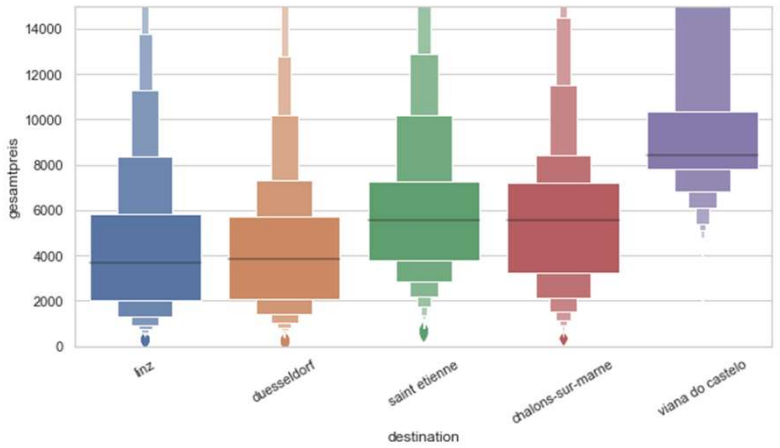


Destinations

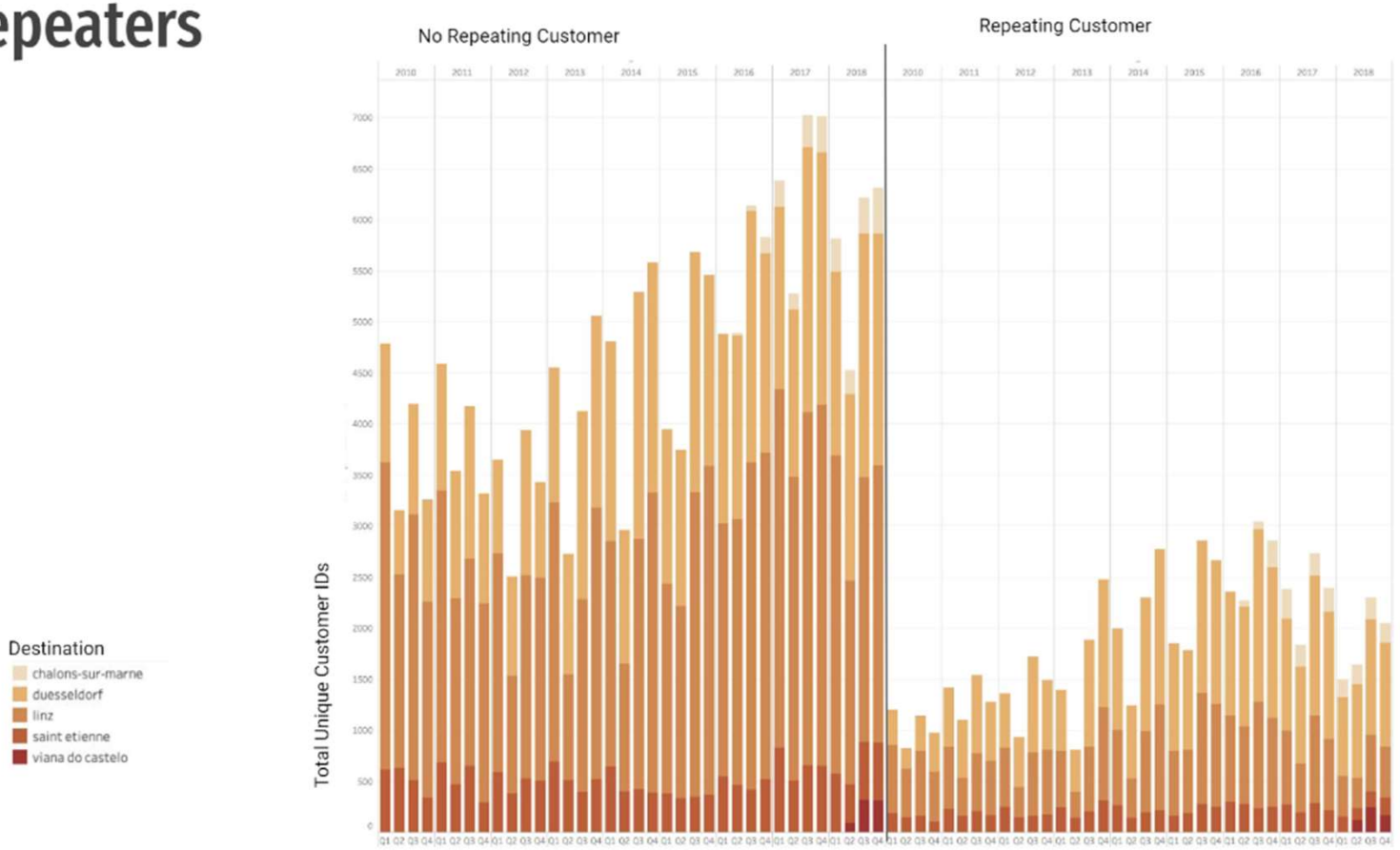
bookings



total amount spent per booking

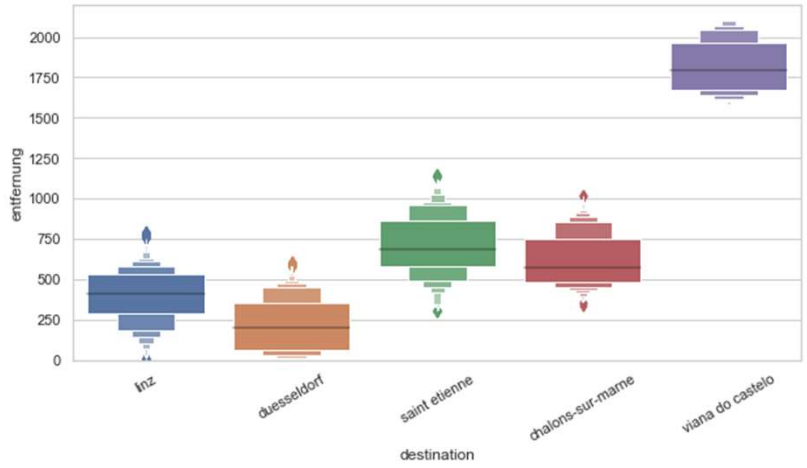


Repeaters

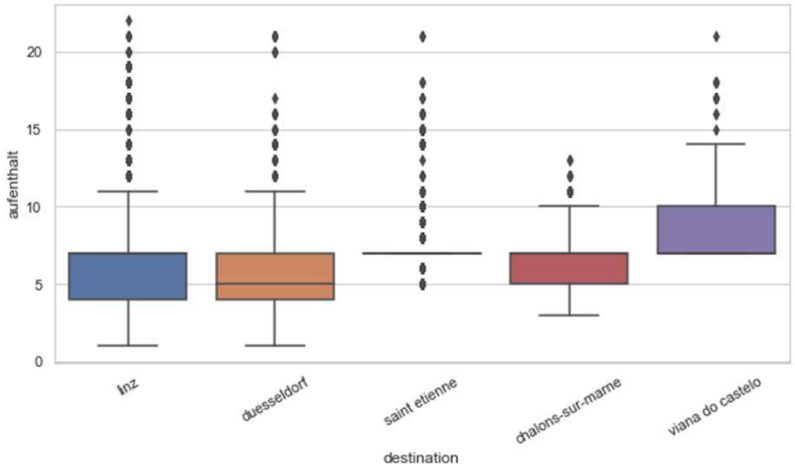

No Repeating Customer
Repeating Customer
Total Unique Customer IDs
Destination
chalons-sur-marne
duesseldorf
linz
saint etienne
viana do castelo

Destinations

distance

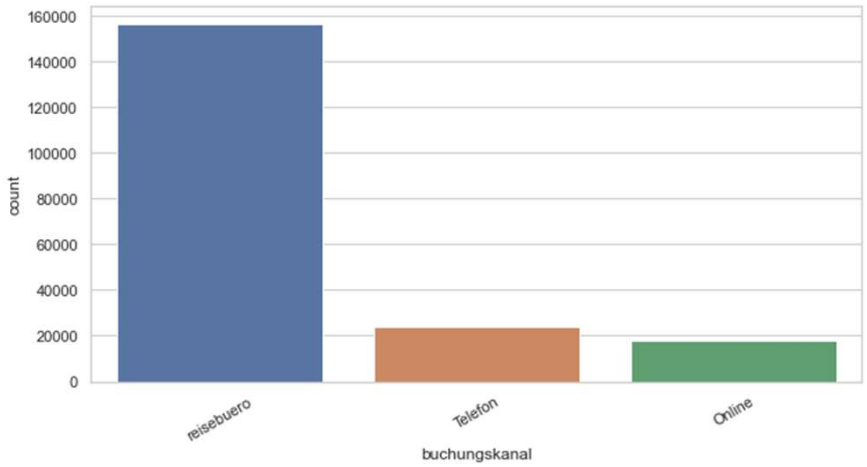


length of stay

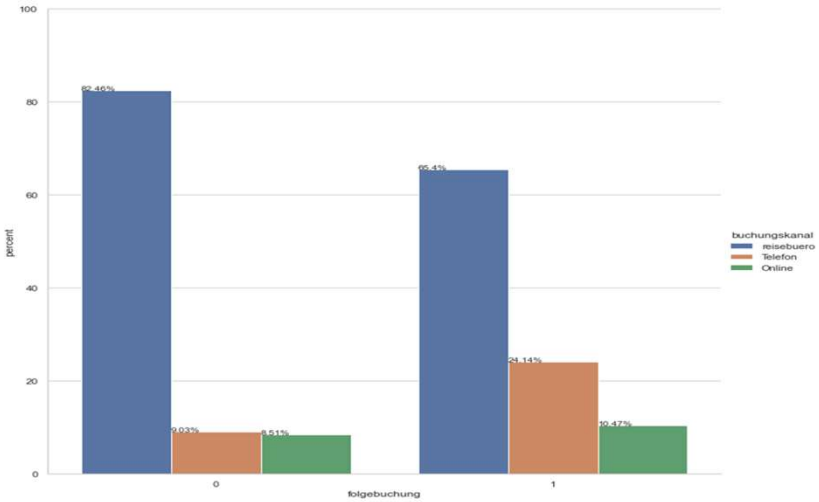


Booking channels

General distribution

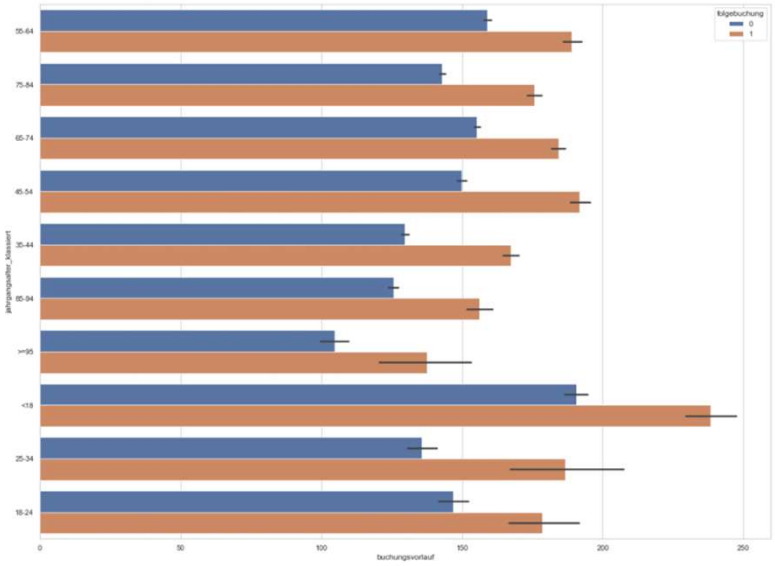


First and follow up bookings

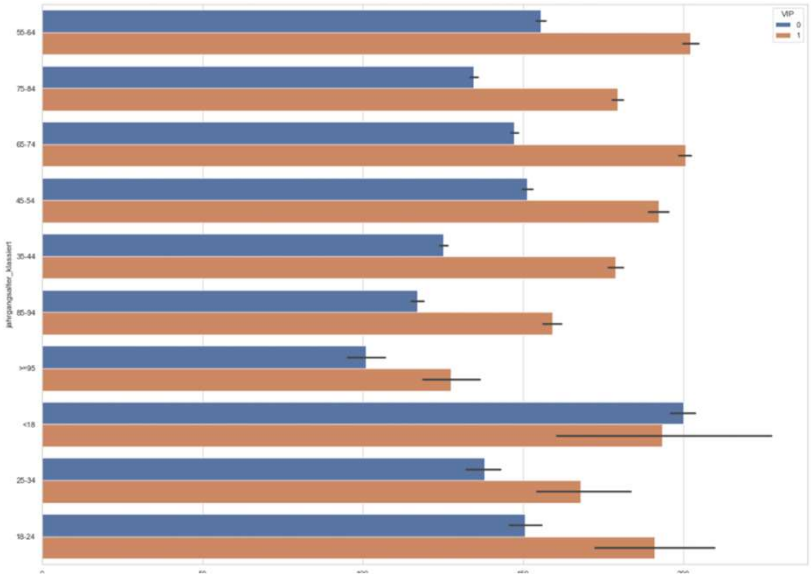


Lead time

by age and repeaters

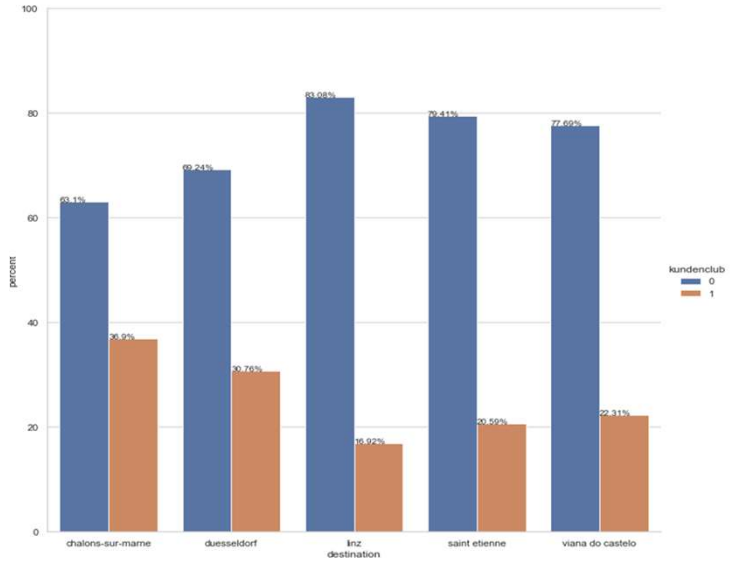


by age and VIPs

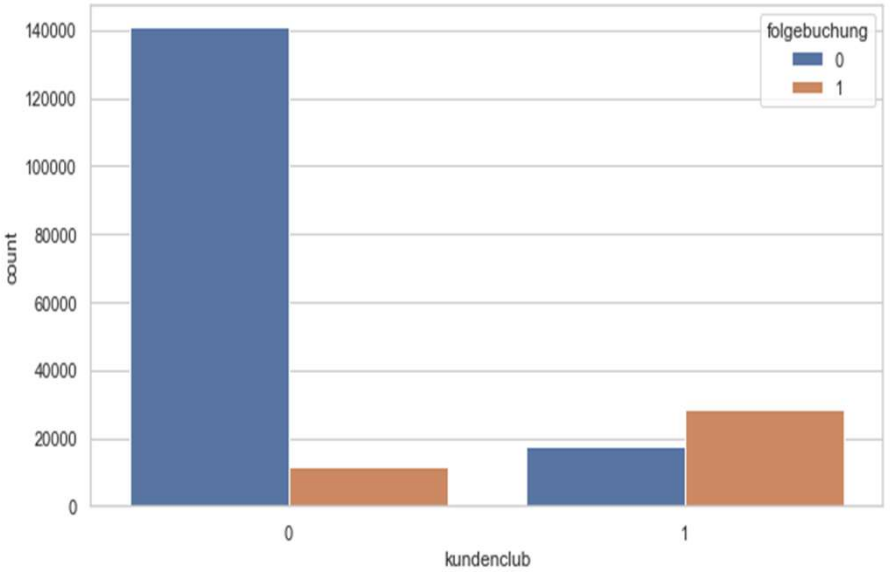


Impact of club membership

Club membership per hotel

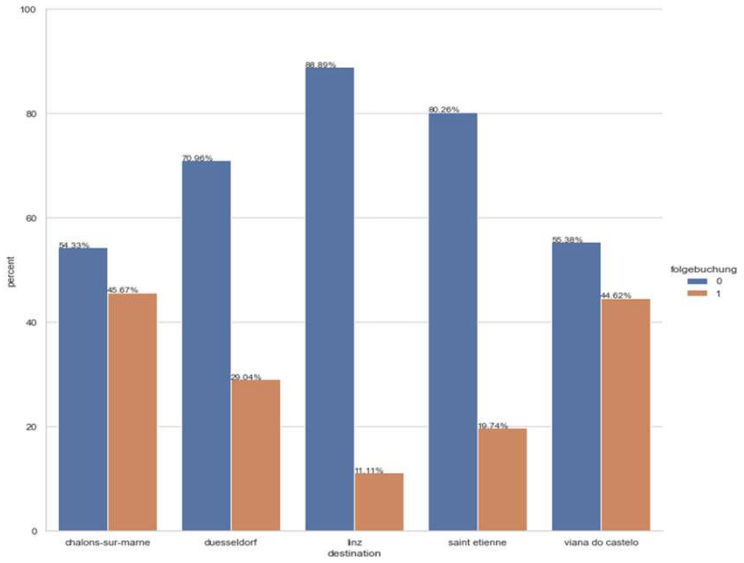


Club member and follow up booking

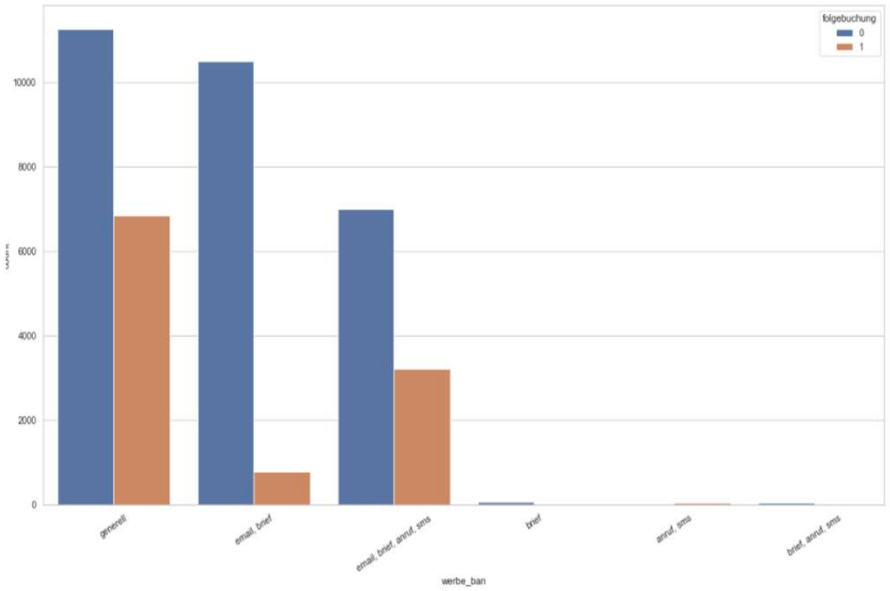


Follow Up bookings

In proportion

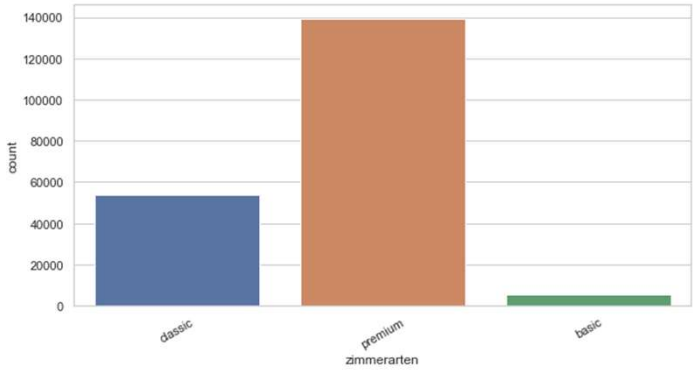


Impact of advertising ban

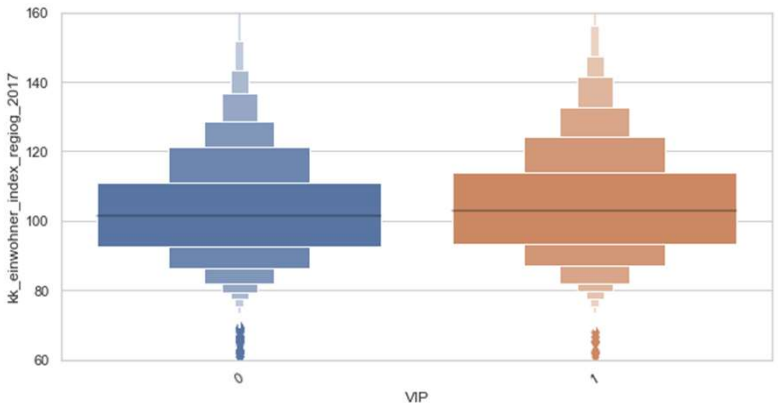


VIPs

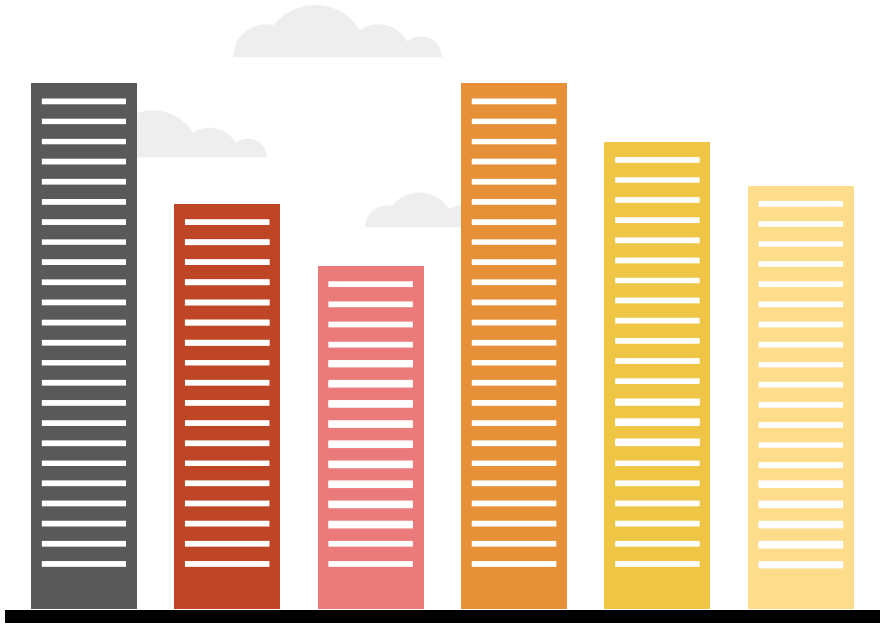
room categories



by purchasing power



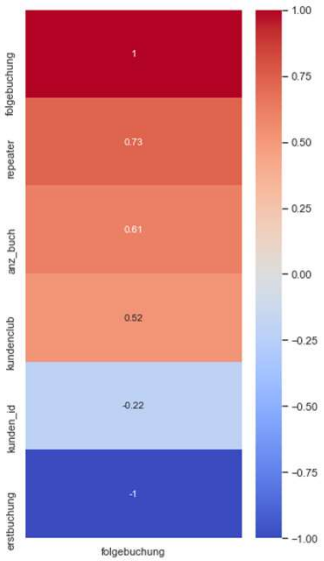
Used modells



- Logistic Regression
- Normalised Logistic Regression
- Logistic Regression with dummies
- Balanced Logistic Regression (up- & downsampled)
- Random Forest
- Gaussian Naive Bayes
- k-Nearest-Neighbor (KNN)
- Decision Trees
- AdaBoost

Best models to predict follow up bookings
(by recall)

Correlation heatmap



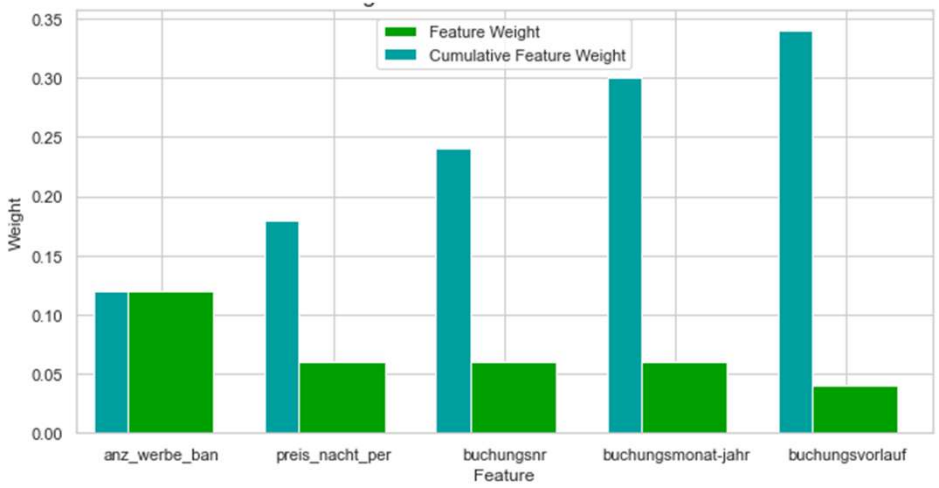
Modell results

- Logistic Regression 5,43%
- Normalised Logistic Regression 18,16%
- Logistic Regression with dummies 26,23%
- Balanced Logistic Regression 63,51% / 59,04%
- Random Forest 46,82%
- GaussianNB 48,26%
- KNN 38,74%
- Decision Trees 52,36%
- AdaBoost 69,74%

Best models to predict follow up bookings

(by recall)

Normalized weights for first five most predictive features

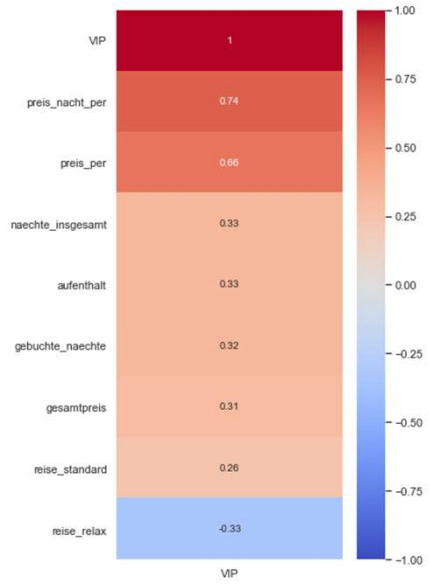


Modell results

- Logistic Regression
  5,43%
- Normalised Logistic Regression 18,16%
- Logistic Regression with dummies 26,23%
- Balanced Logistic Regression 63,51% / 59,04%
- Random Forest
  46,82%
- GaussianNB
  48,26%
- KNN 38,74%
- Decision Trees
  52,36%
- AdaBoost
  69,74%

Best models to predict VIPs
(by recall)

Correlation heatmap



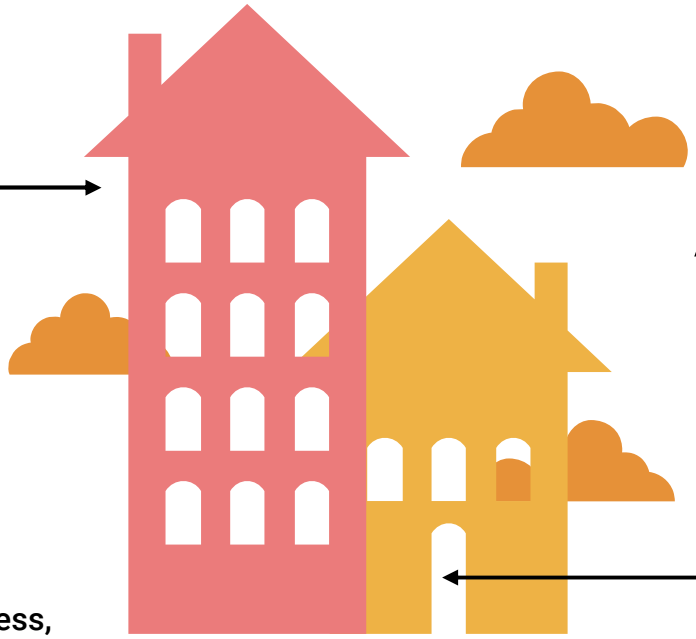
Modell results

- Logistic Regression 32,77%
- Normalised Logistic Regression 45,87%
- Logistic Regression with dummies 68,47%
- Balanced Logistic Regression 79,13% / 65,81%
- Random Forest 45,40%
- GaussianNB 60,44%
- KNN 36,89%
- Decision Trees 47,68%
- AdaBoost 47,61%

Outlook

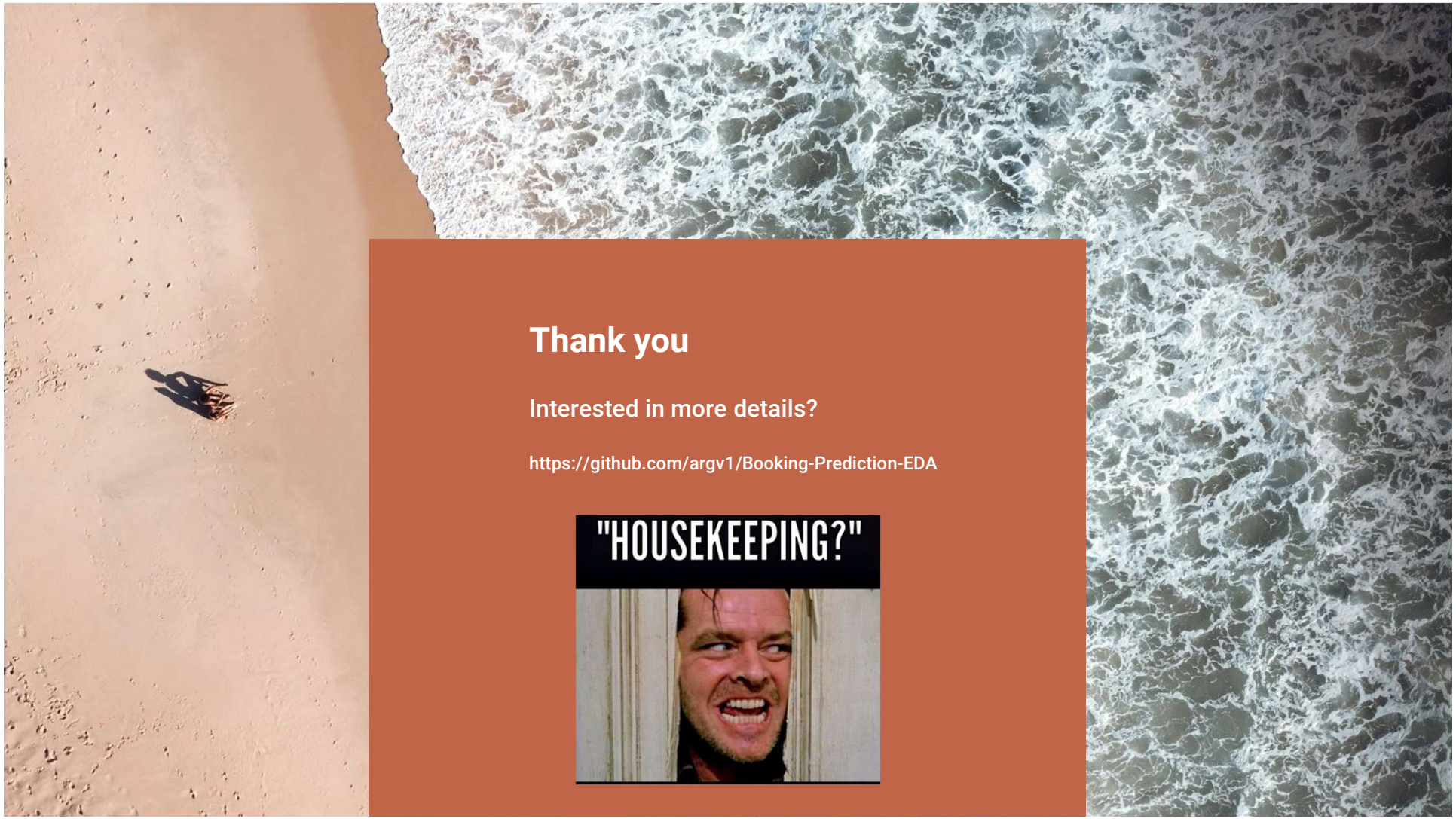
Adding features

- Events
- Customer feedback
- Cancellations
- Denails
- Overbooking strategy
- Revenue per room
- Hotel occupancy rate
- Booking Category (business, leisure, airline, fair, etc.)



Answer more research questions

- Forecast rates
- Predict demand
- Estimate risk of cancellation
- Recommend overbooking



Thank you

Interested in more details?

https://github.com/argv1/Booking-Prediction-EDA