HOTEL BOOKINGS: „TO CANCEL OR NOT TO CANCEL“

DIANA JAFFÉ
Code Academy
June 11,2021

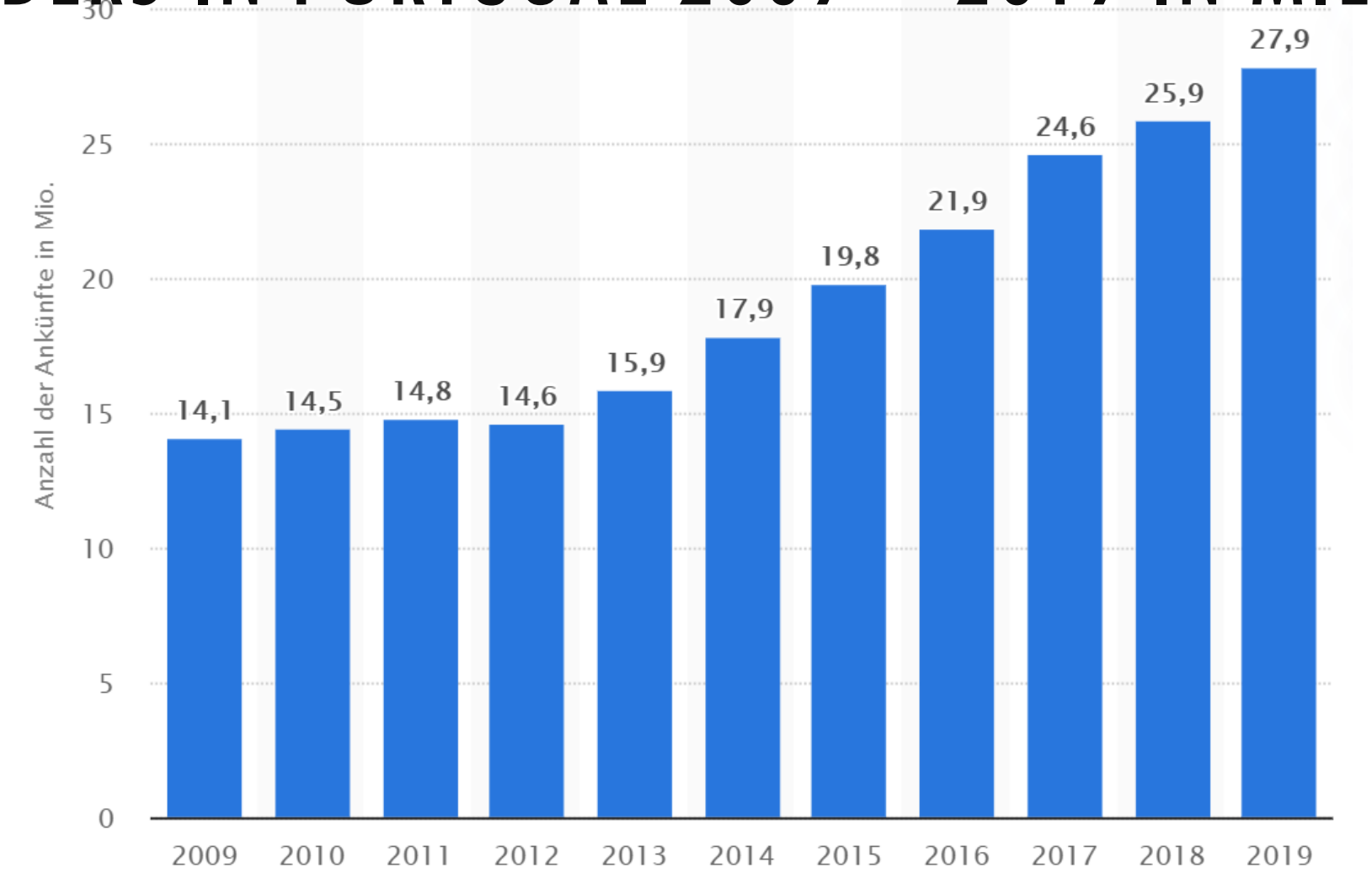
HOTELS: „TO CANCEL OR NOT TO CANCEL“

- Where does the data come from?
- A tiny little bit about the Portugese travel market.
- Hidden secrets in the data.
- Machine Learning: Can „the machine“ foretell people‘s behavior?

THE PORTUGESE TOURISTIC MARKET - 2018

- Tourism made 8.52% of the Portugese GDP in 2018.
- International travelers make 71% of the 57.6m overnight stays *in hotels*.
- This equals 12.7m arrivals of international guests.
- Overnight stays *in hotels* have dropped in a range from 1% - 5% since 2016.
- RevPAR has nevertheless increased.

TOURIST ARRIVALS AT ACCOMODATION PROVIDERS IN PORTUGAL 2009 – 2019 IN MILL.



MARKET INFORMATION

Tourism Demand

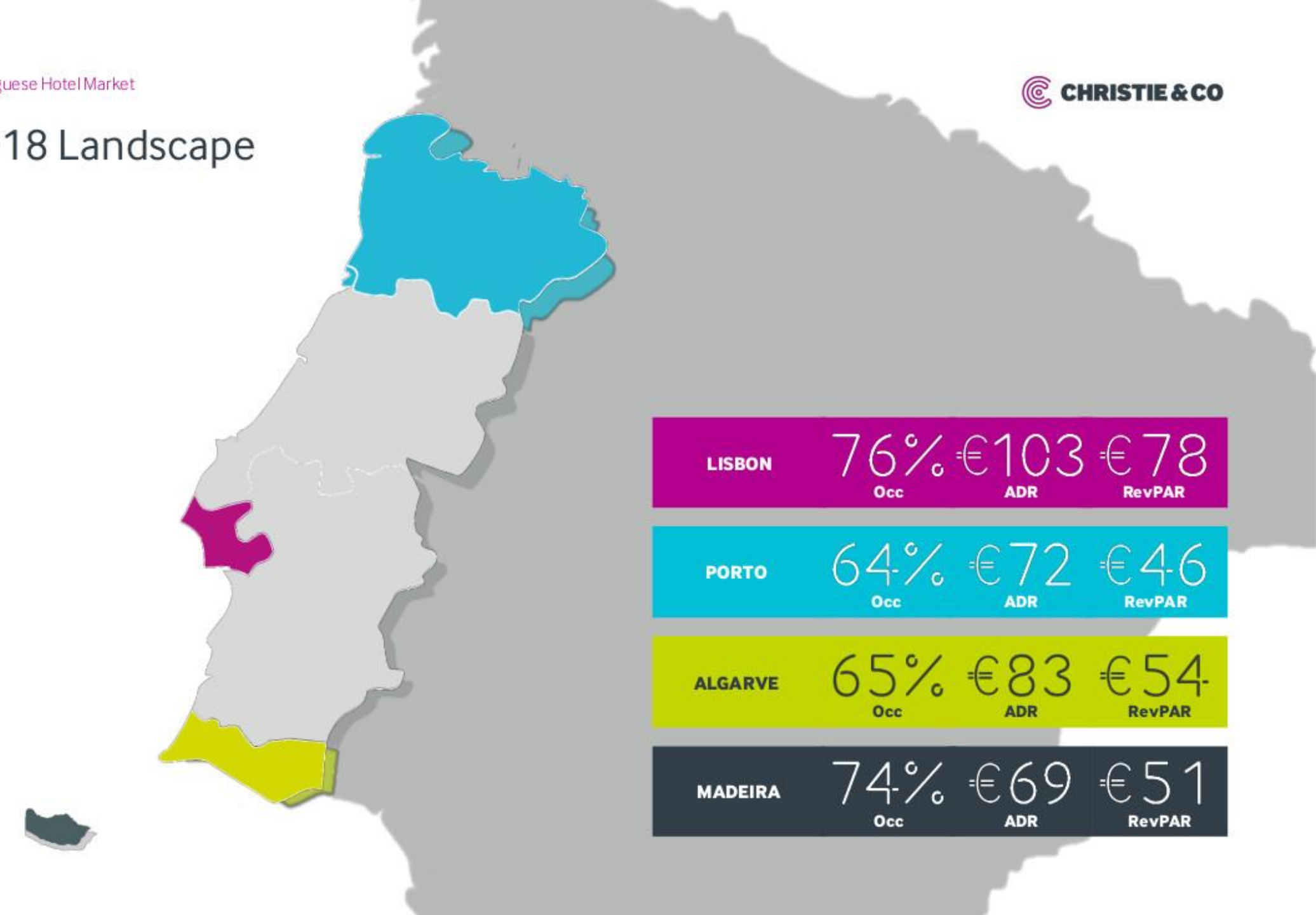
Hotel Supply

Hotel Performance

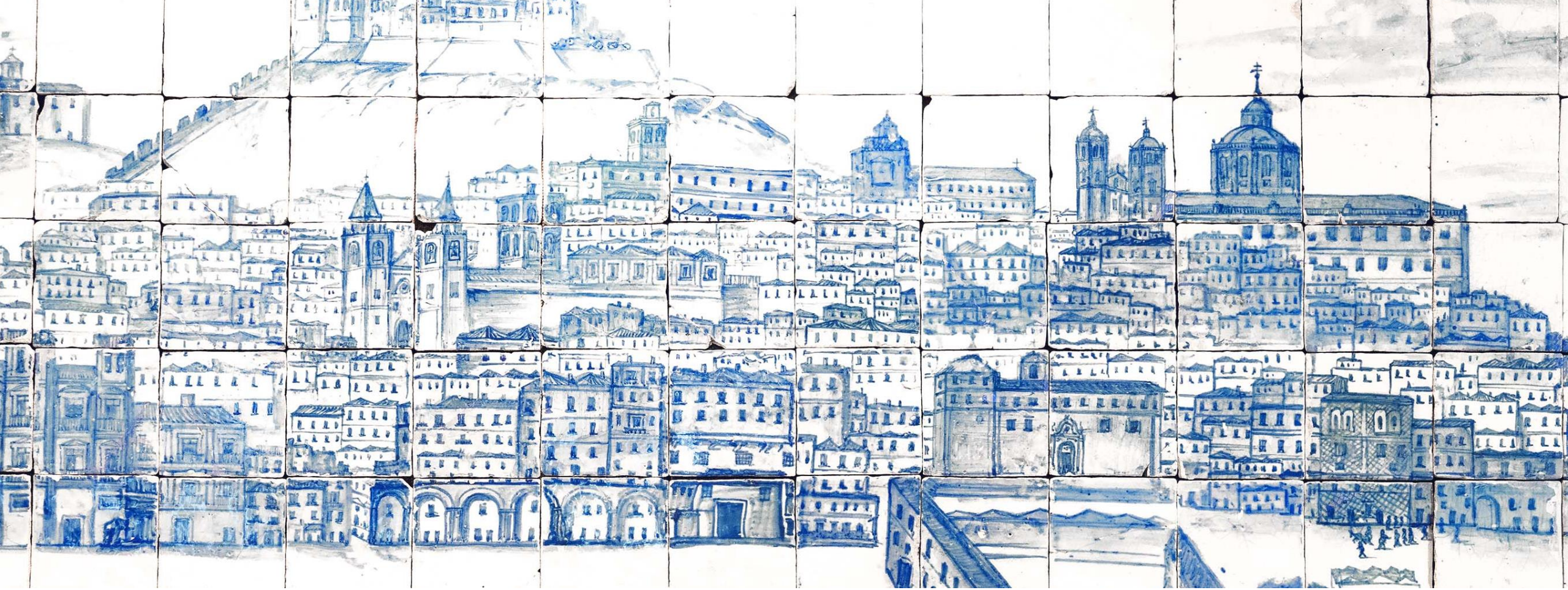
Portuguese Hotel Market

2018 Landscape

CHRISTIE & CO



	Occ	ADR	RevPAR
LISBON	76%	€103	€78
PORTO	64%	€72	€46
ALGARVE	65%	€83	€54
MADEIRA	74%	€69	€51



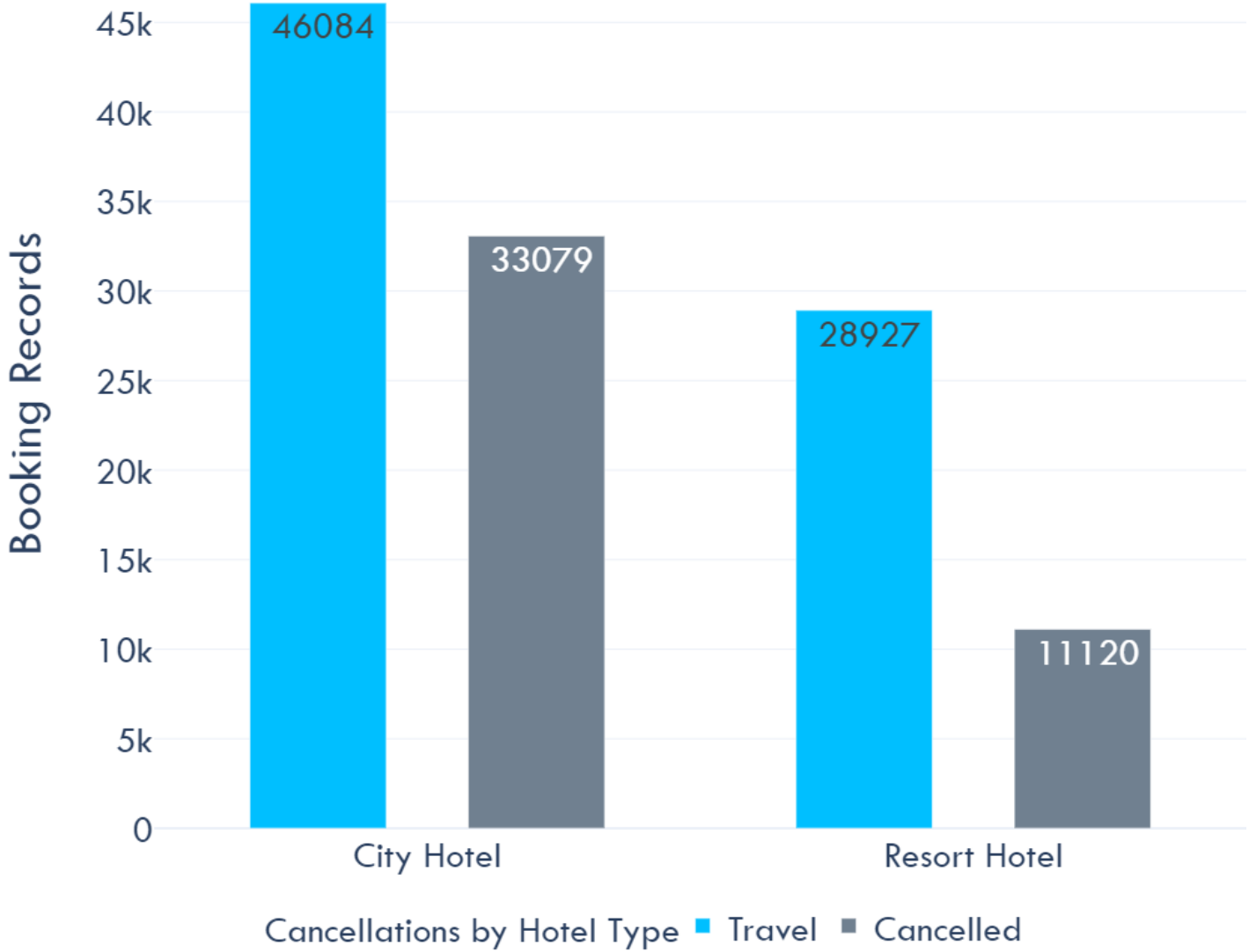
THE BEAUTY - OF DATA

Good Data Paints A Picture Of The World.

KPI

- Everything is interesting.
- Especially people‘s behavior.

DATA STRUCTURE



FINDINGS

- The guests of the 2 hotels originate from 178 countries!

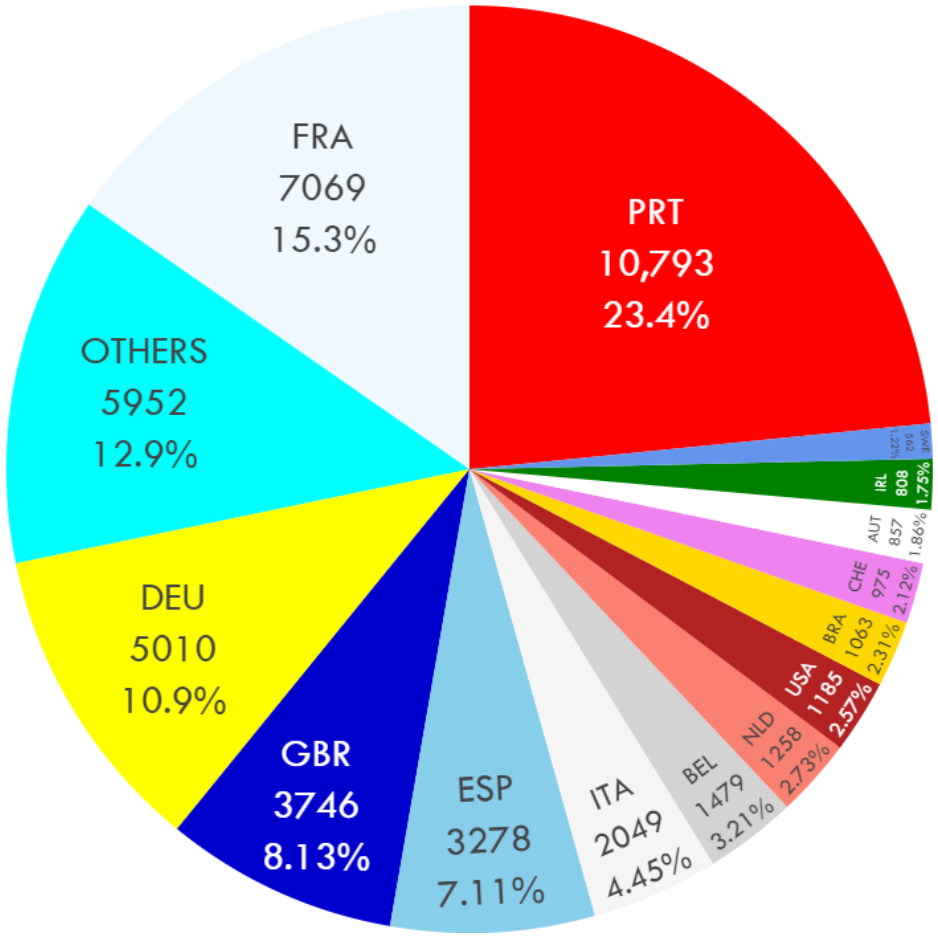
- It‘s surprising how buggy many published data explorations and machine learning codes are!
  And some of the publishers look so professional that you would not doubt what they write – unless you want to work with their codes. So beware!
- No matter how long you work with the data – senseless or corrupt data keeps appearing all the time. E. g.:
  - 251 records for City Hotels are neither canceled, nor do they have overnight stays; the price (adr) == 0
  - Same is true for 371 bookings in the category Resort Hotel.

MEAN ADR CHANGES BY SEASON / MONTH

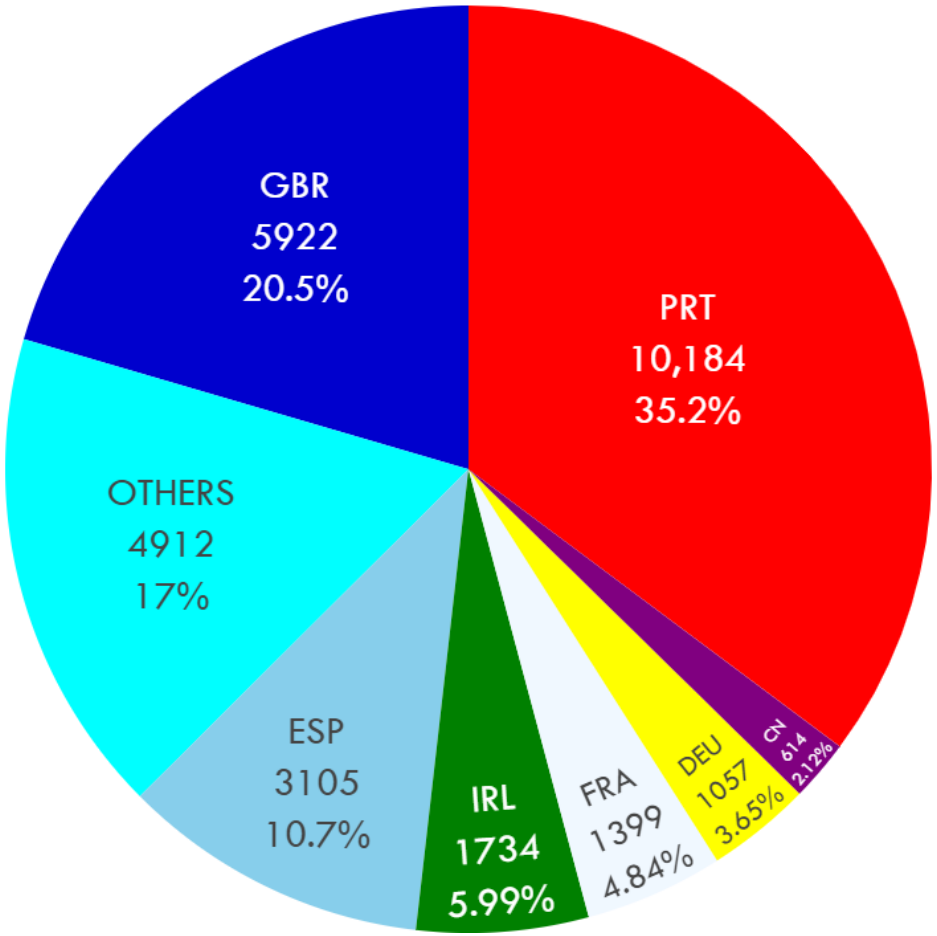
ADR BY ADULTS ONLY / FAMILIES (TRAVEL/CANCELLATION)

BOOKINGS BY COUNTRY AND HOTEL TYPE

Bookings City Hotel

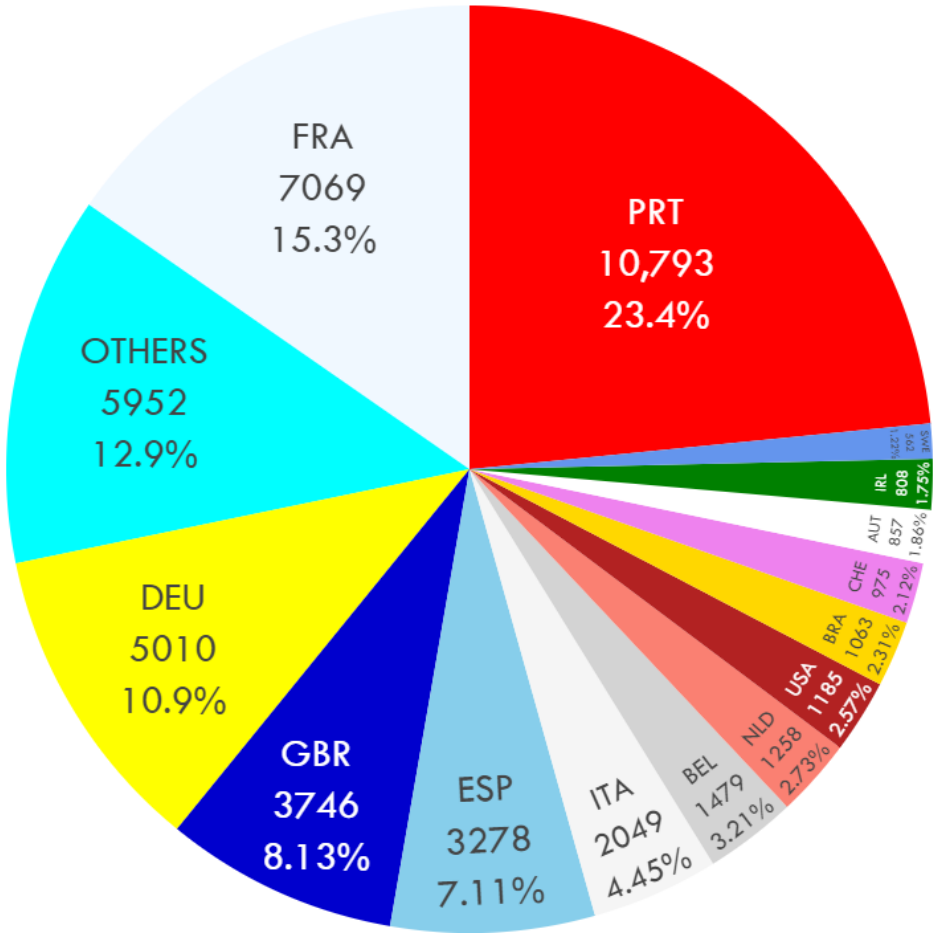


Bookings Resort Hotel

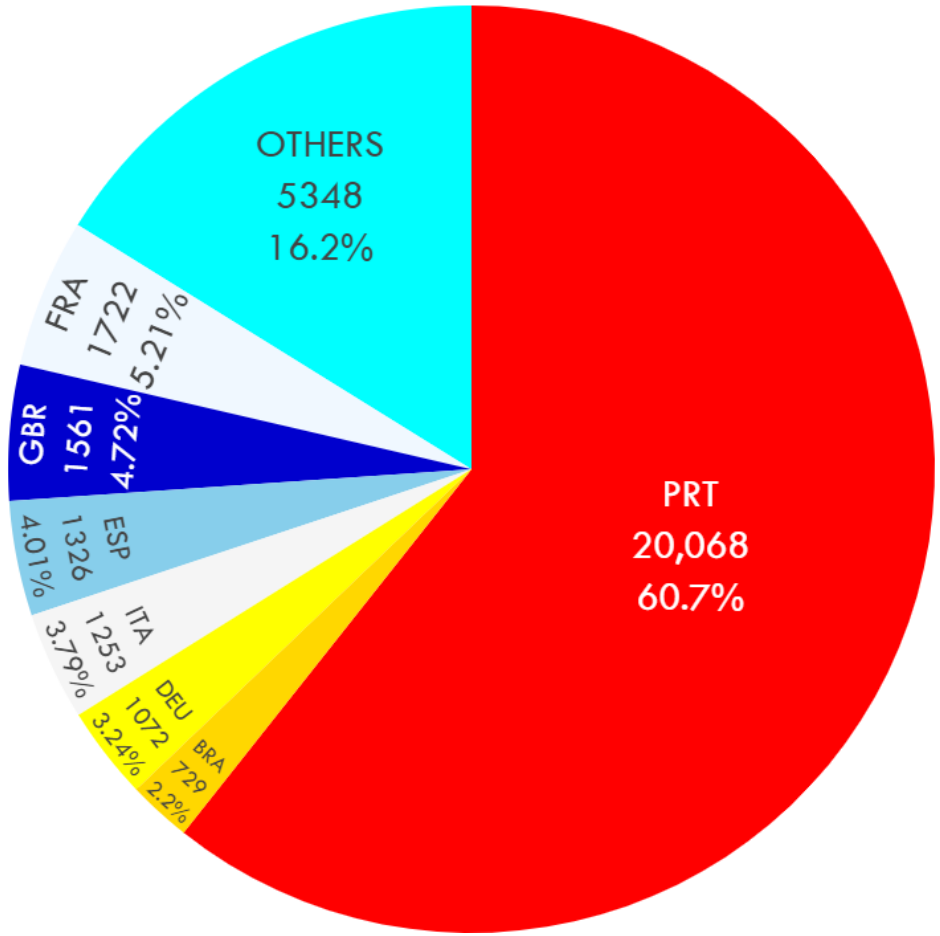


BOOKINGS AND CANCELLATIONS BY COUNTRY

Bookings City Hotel

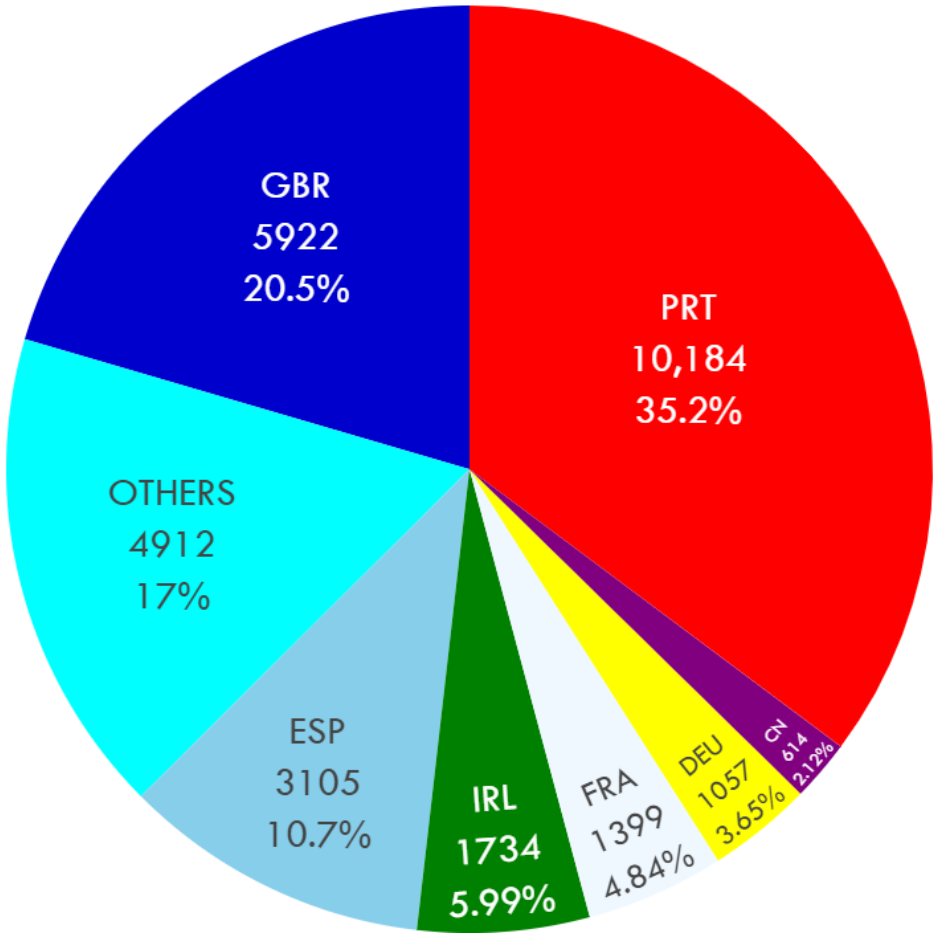


Cancellations City Hotel

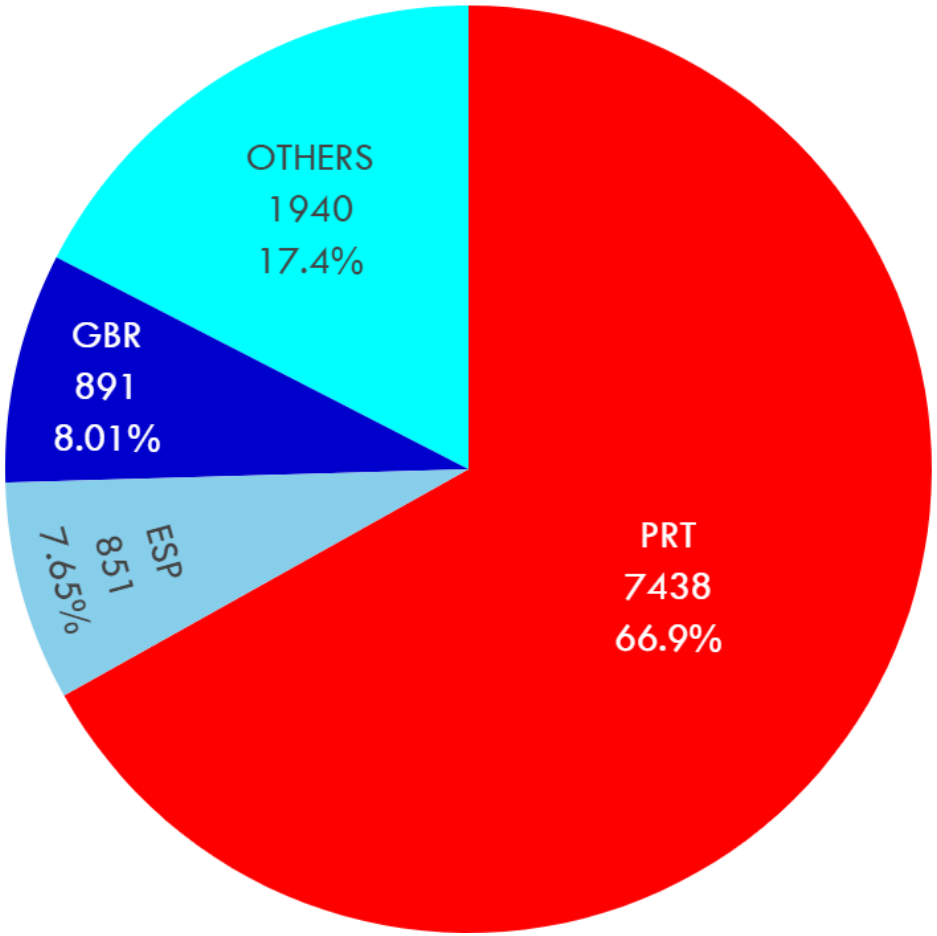


BOOKINGS AND CANCELLATIONS BY COUNTRY

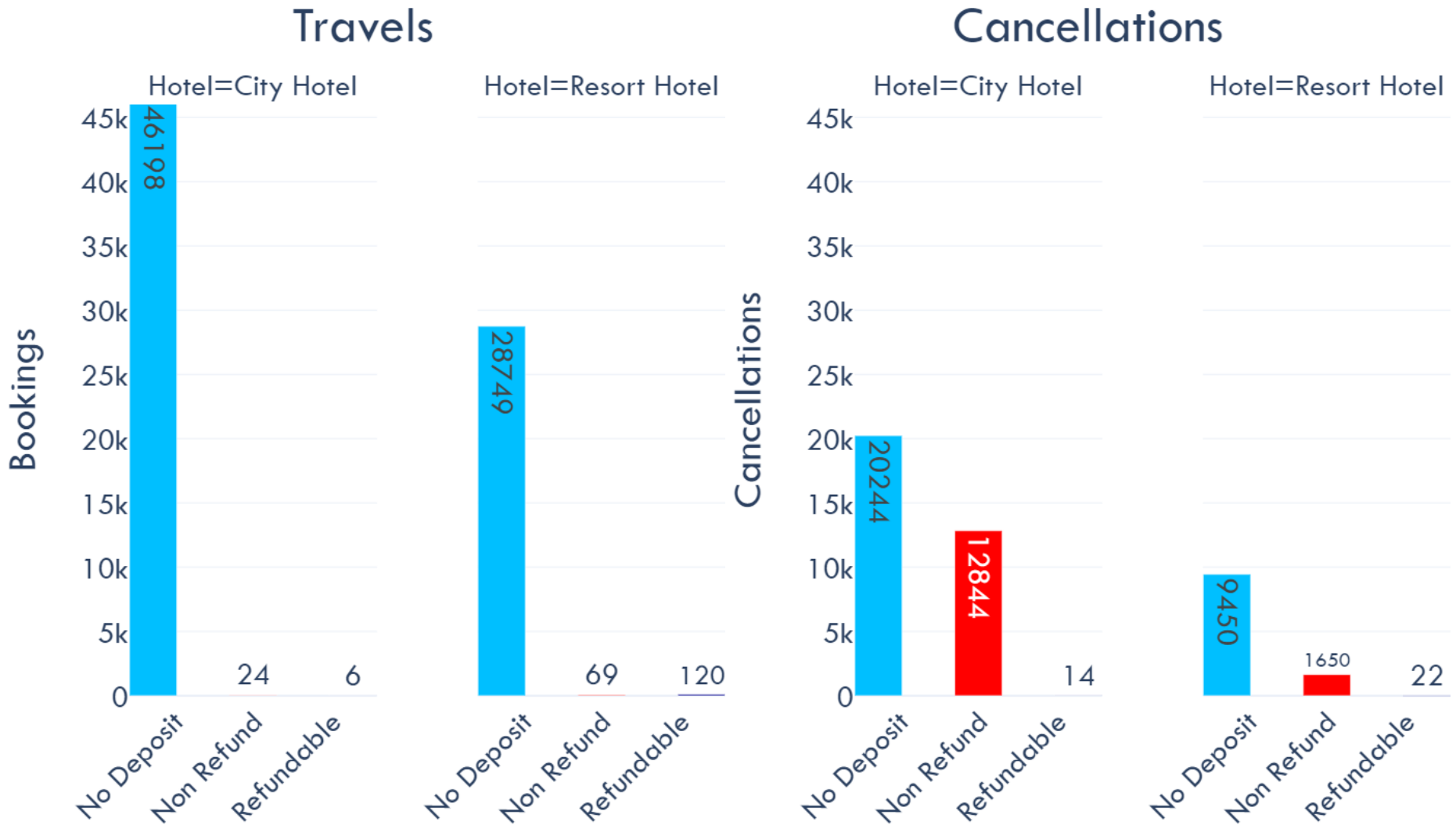
Bookings Resort Hotel



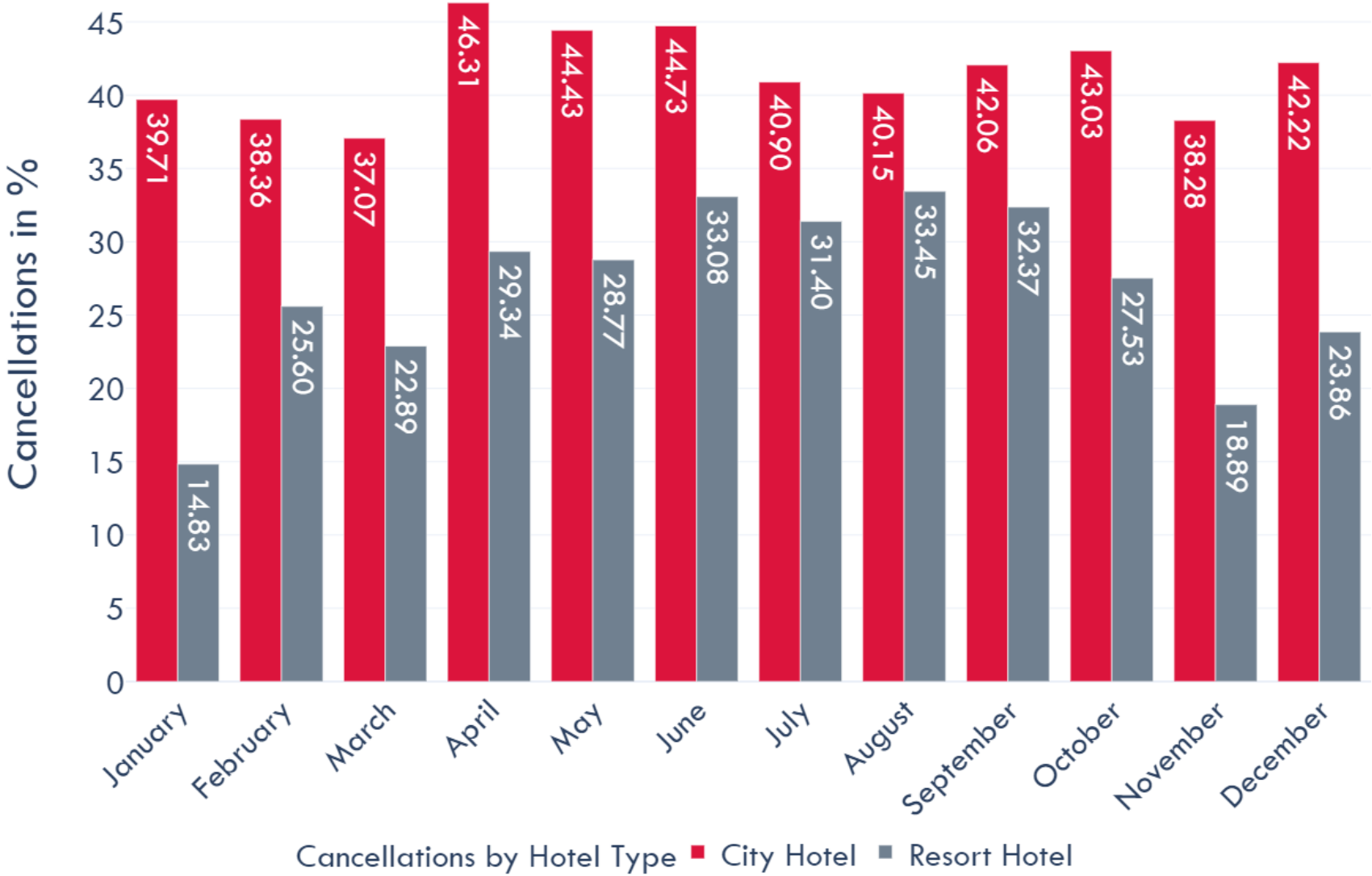
Cancellations Resort Hotel



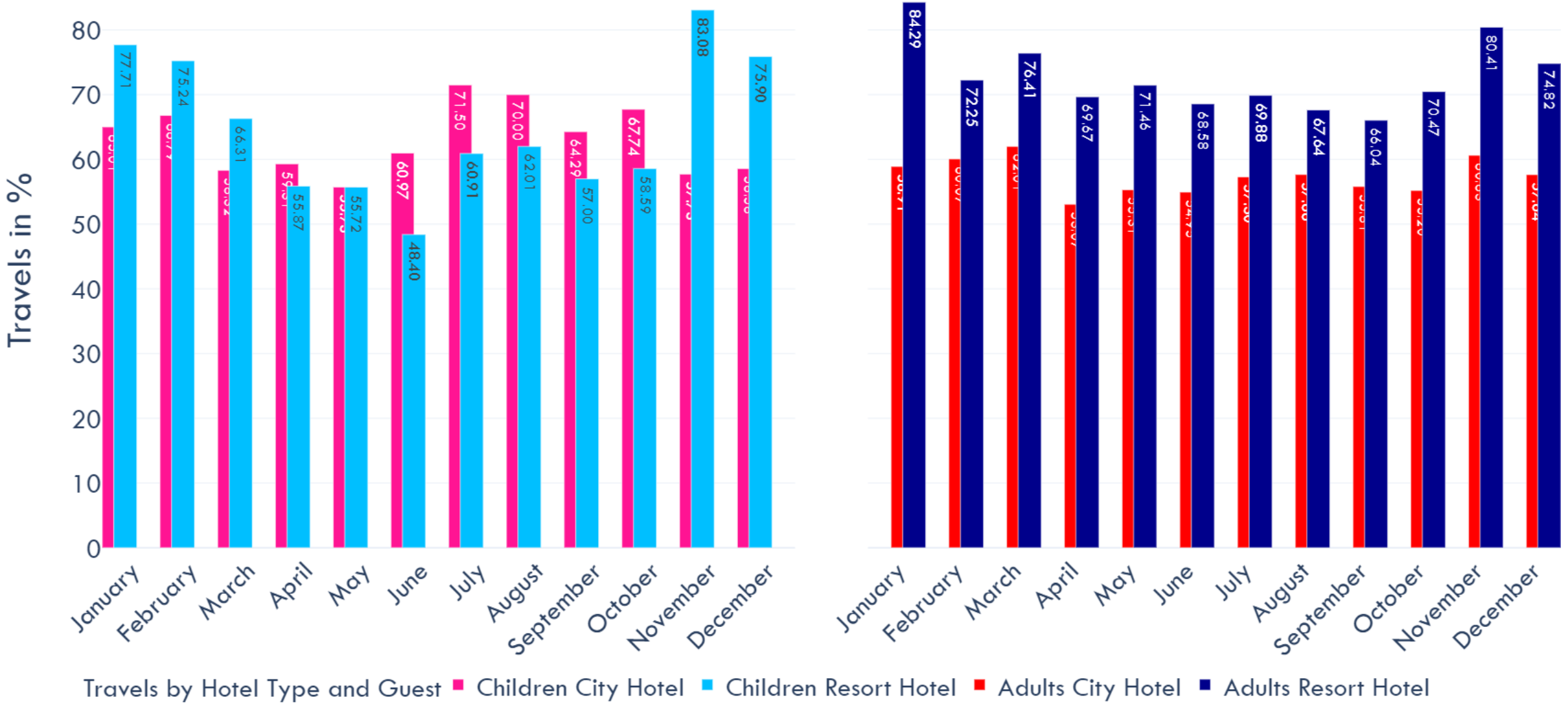
BOOKINGS AND CANCELLATIONS BY DEPOSIT



CANCELLATION RATE PER MONTH IN %

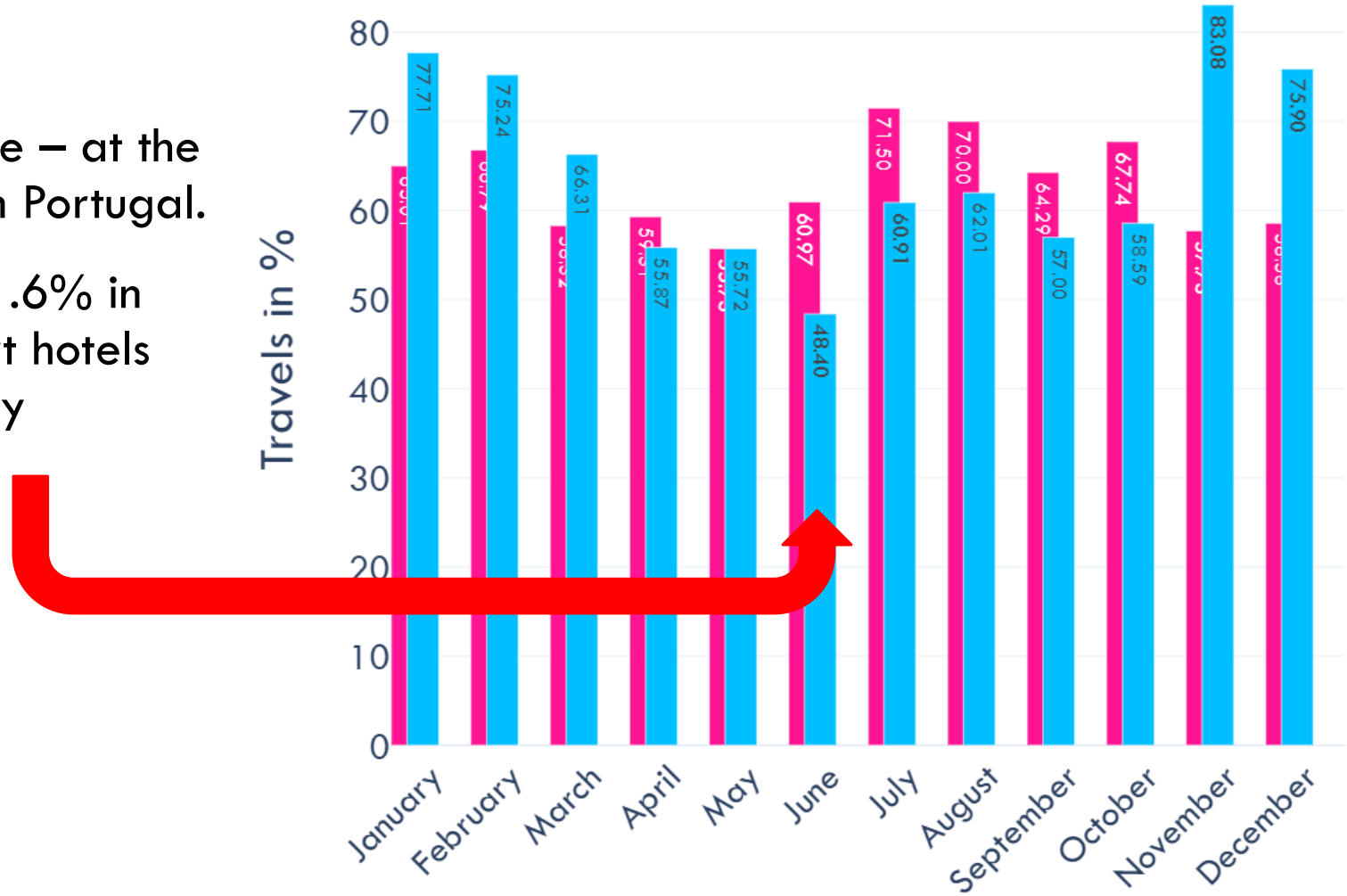


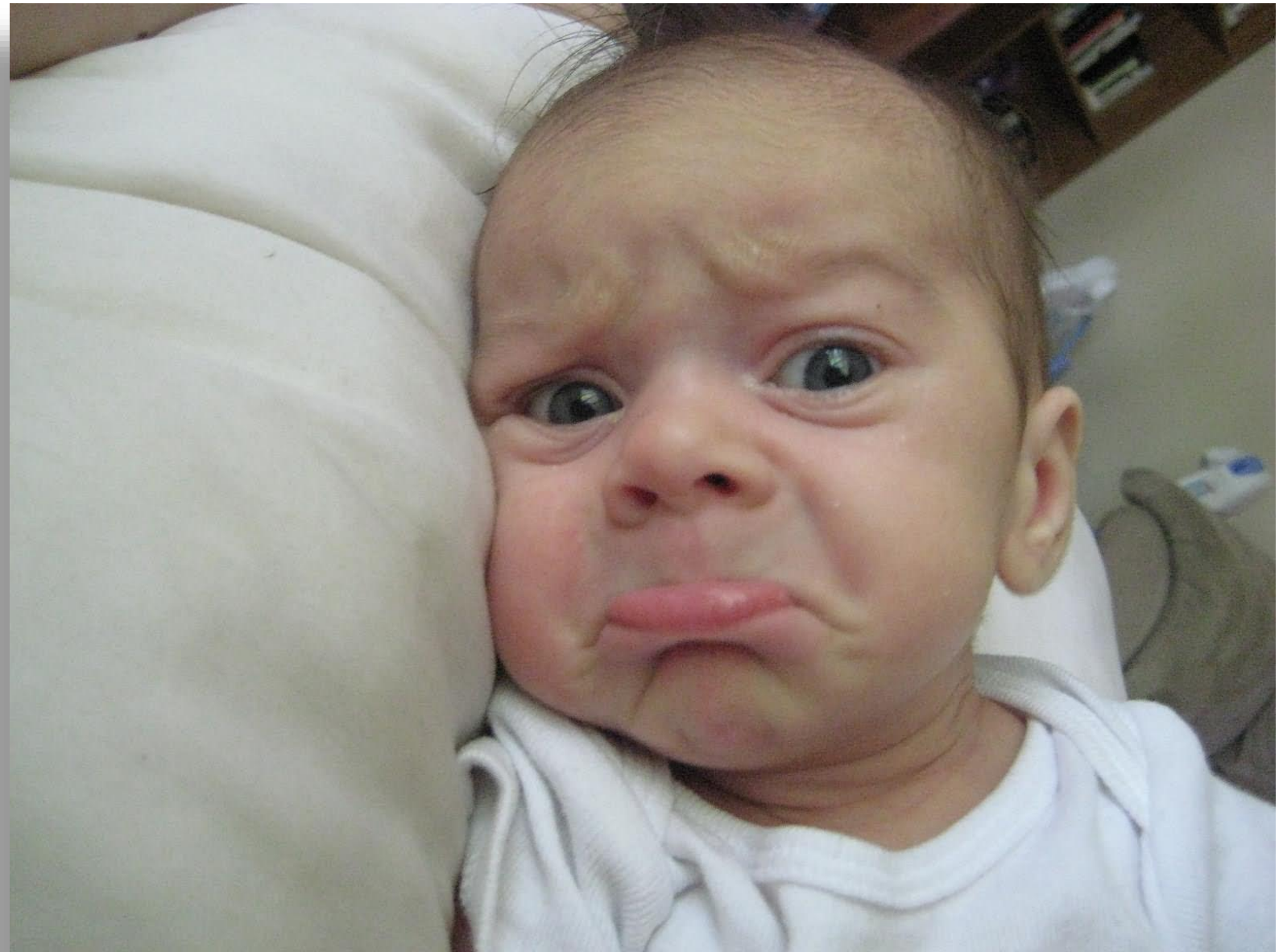
TRAVEL RATE (POSITIVE FOR NO CANCELLATION)



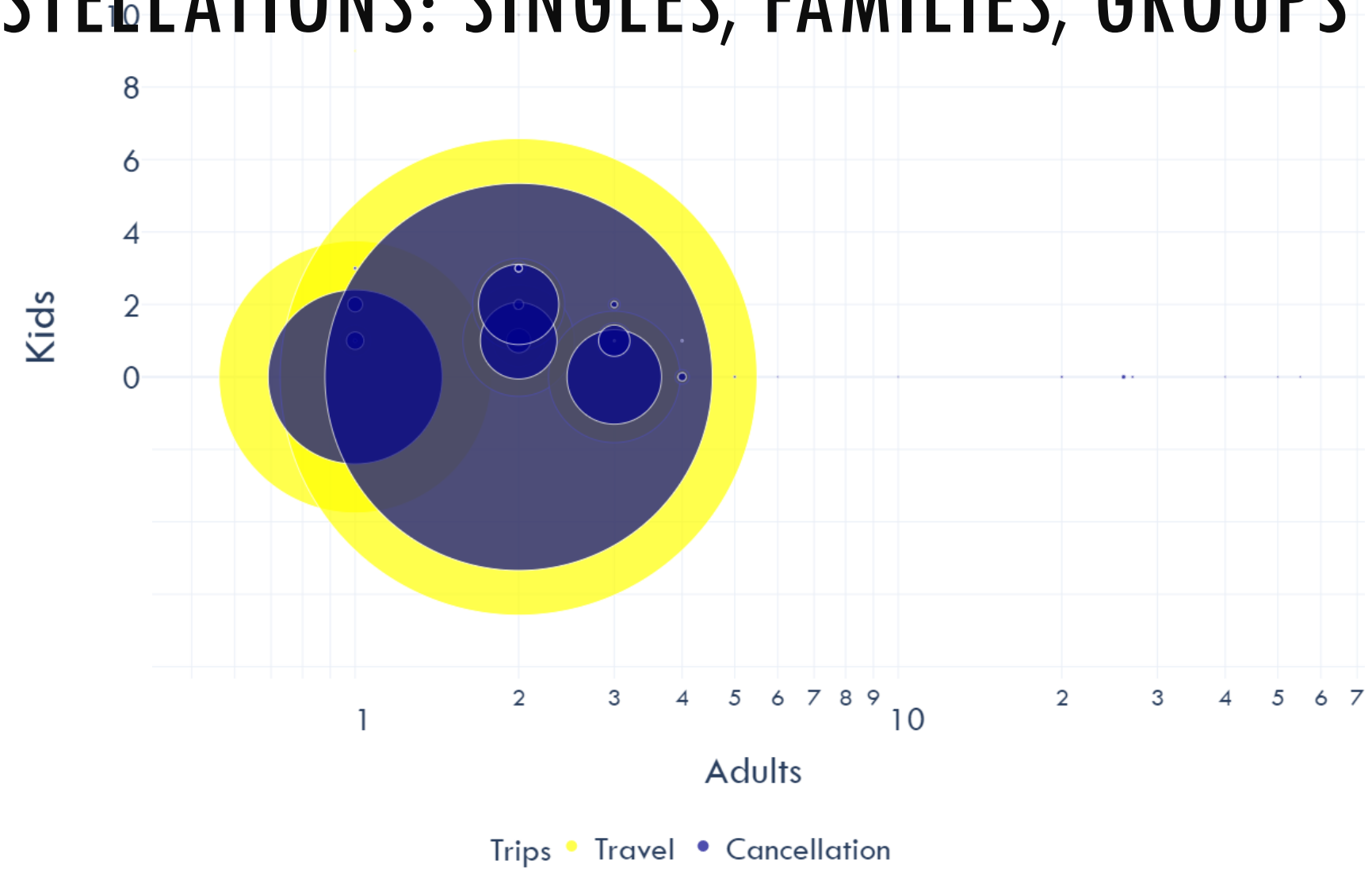
MOST CANCELLATIONS AFFECT CHILDREN

- Most cancellations occur in June – at the start of the summer holidays in Portugal.
- With a cancellation rate of 51.6% in families with children for resort hotels children are overproportionally affected.





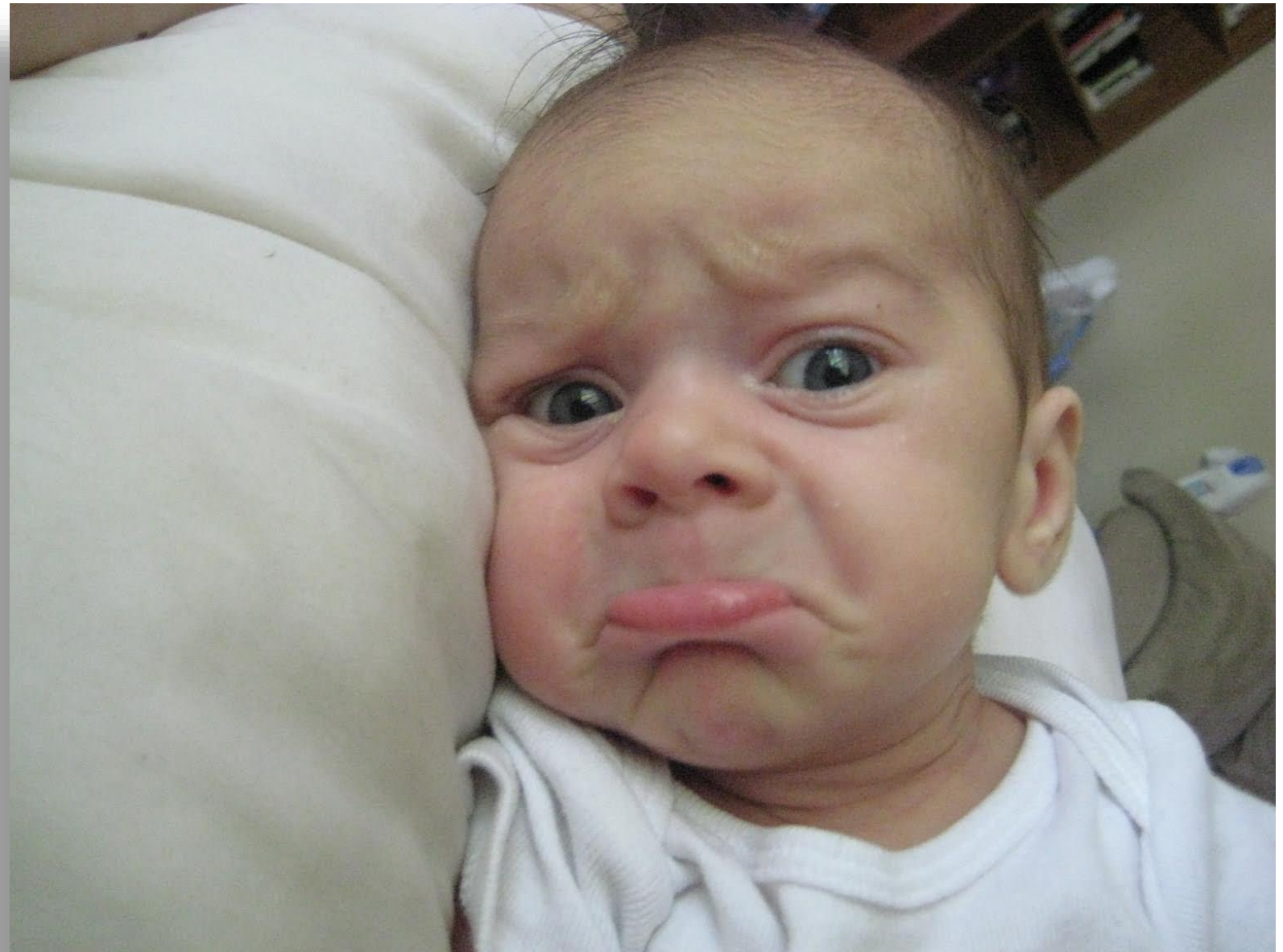
TRAVEL AND CANCELLATION RATES BY TRAVEL CONSTELLATIONS: SINGLES, FAMILIES, GROUPS ETC.



LIFE (AND BUSINESS) BEGINS WHERE DATA ENDS

Questions:

- [How can businesses prevent senseless or corrupt data in their systems?
  Or is that just part of a very special information gathering?]
- What are the reasons for the vast cancellation rates for resort hotels in summer of families with children?
- Can hotel owners can do anything to change this (not only for their businesses, but also for families‘ sake)?
- Are resort hotels always fully booked in summer so they don‘t have to care about cancellations? – We did not get any occupation rates for the two hotels.
- But not everything is always just business…



CAN A MACHINE PREDICT PEOPLE'S BEHAVIOR IN THE FUTURE?

TRYING TO TACKLE THE CANCELLATION QUESTION

	MODEL	SCORE	KAPPA
1	Random Forest Classifier	0.891871	0.765818
2	Decision Tree Classifier	0.855423	0.691746
3	KNN	0.835123	0.643846
4	K-Fold Decision Tree Classifier	0.819173	
5	Logistic Regression	0.792928	0.529230

„NICE. BUT YOU CAN DO BETTER.“

Accuracy > 90 %
Kappa > 80%

AND I TRIED HARDER...

Strategies

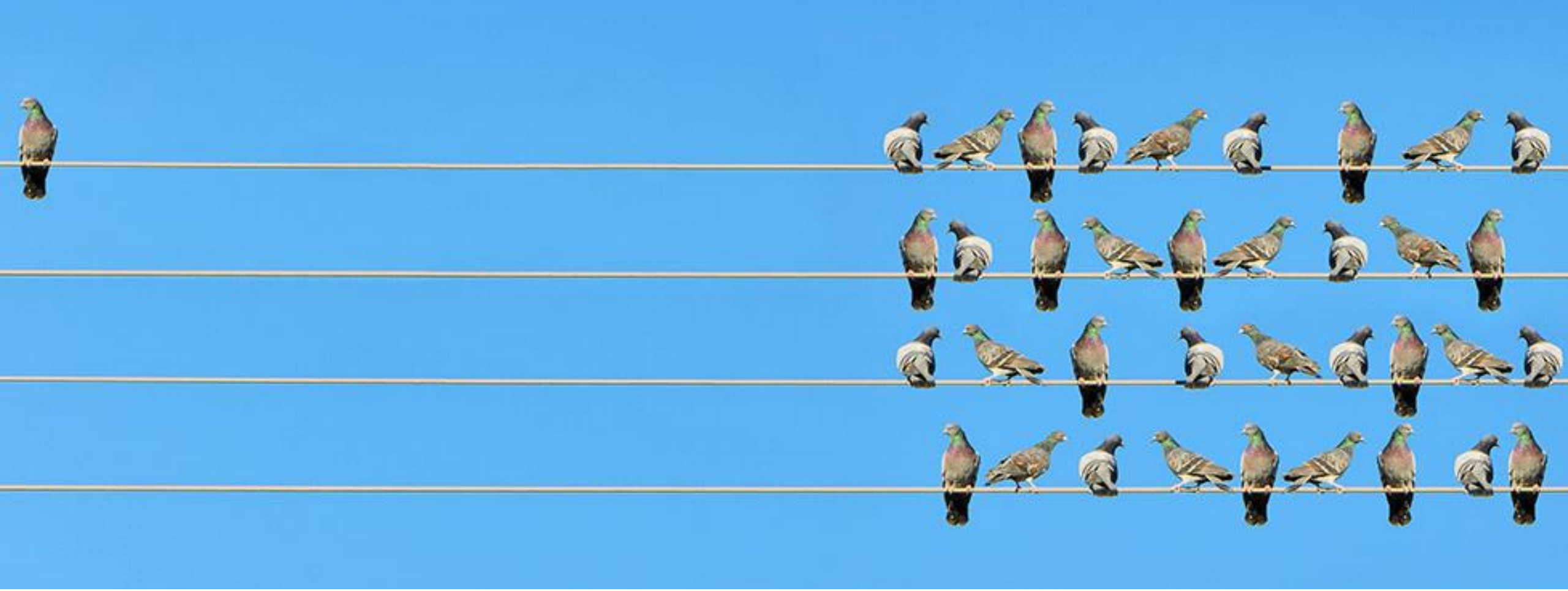
- Standardization
- Normalization
- Dummification
- Outlier deletion
- Correlation Matrix
- Feature selection
- Splitting hotel data
- Different random states.

Classification Methods

- Logistic Regression
- Decision Tree
- K-Fold Decision Tree
- KNN
- Random Forest
- Ada Boost Classifier
- Gradient Boosting Classifier
- XgBoost Classifier
- Cat Boost Classifier
- Extra Trees Classifier
- LGBM Classifier
- Voting Classifier.

TRYING TO TACKLE THE CANCELLATION QUESTION

	MODEL	SCORE	KAPPA
1	Random Forest Classifier	0.894053	0.770164
2	Extra Trees Classifier	0.889774	0.759985
3	Voting Classifier	0.879414	0.735213
4	Cat Boost	0.876730	0.733468
5	XgBoost	0.868258	0.714031
6	LGBM	0.863476	0.706467
7	Decision Tree Classifier	0.856094	0.693108
8	Gradient Boosting Classifier	0.849425	0.668910
9	KNN	0.835794	0.646068
10	Ada Boost Classifier	0.838520	0.644882
11	Logistic Regression	0.794313	0.535256



THE PROBLEM: IMBALANCED DATA

THE SOLUTION: OVERSAMPLING WITH SMOTE

FINALLY THE MIRACLE…

	MODEL	SCORE	KAPPA
1	Random Forest Classifier	0.931811	0.863627
2	Extra Trees Classifier	0.930645	0.861290
3	K-Fold Decision Tree	0.811953	0.811953
4	Decision Tree Classifier	0.904816	0.809642
5	Voting Classifier	0.894484	0.788964
6	Cat Boost	0.876487	0.752982
7	XgBoost	0.870488	0.740984
8	LGBM	0.866389	0.732787
9	KNN	0.842260	0.684540
10	Gradient Boosting Classifier	0.840927	0.681851
11	Ada Boost Classifier	0.839293	0.678582

THANK YOU

DIANA JAFFÉ
Code Academy
June 11,2021