Exploring **Hotel Demand** through Data Science

About the Dataset

Hotel Booking Transactions

1 City Hotel & 1 Resort Hotel

Located in Portugal

Located in region with high hospitality competition

31 Variables & over 100k Rows

Spanning 2 years

Shared Business Owner

Business Problem

What?

Increase overall Year-on-Year revenue for hotel.

Why?

Improvement project in hotel franchise.

Who?

A group of data analysts hired by hotel owner.

How?

Given a **starting budget** of $150,000.

Problem Evaluation

Overbooking

- What is Overbooking?
- What are the benefits of overbooking?
- How do we optimise the overbooking strategy?
- How should we do when customers are affected?

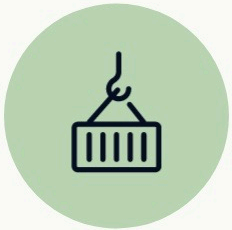
Overbooking in Real World

Passenger Airlines

Hotels
(Business & Resort)

Restaurants

Car Rentals



Cargo Freight

Health Clinics

Cruise Lines

Events
(E.g. Sports, Concerts)

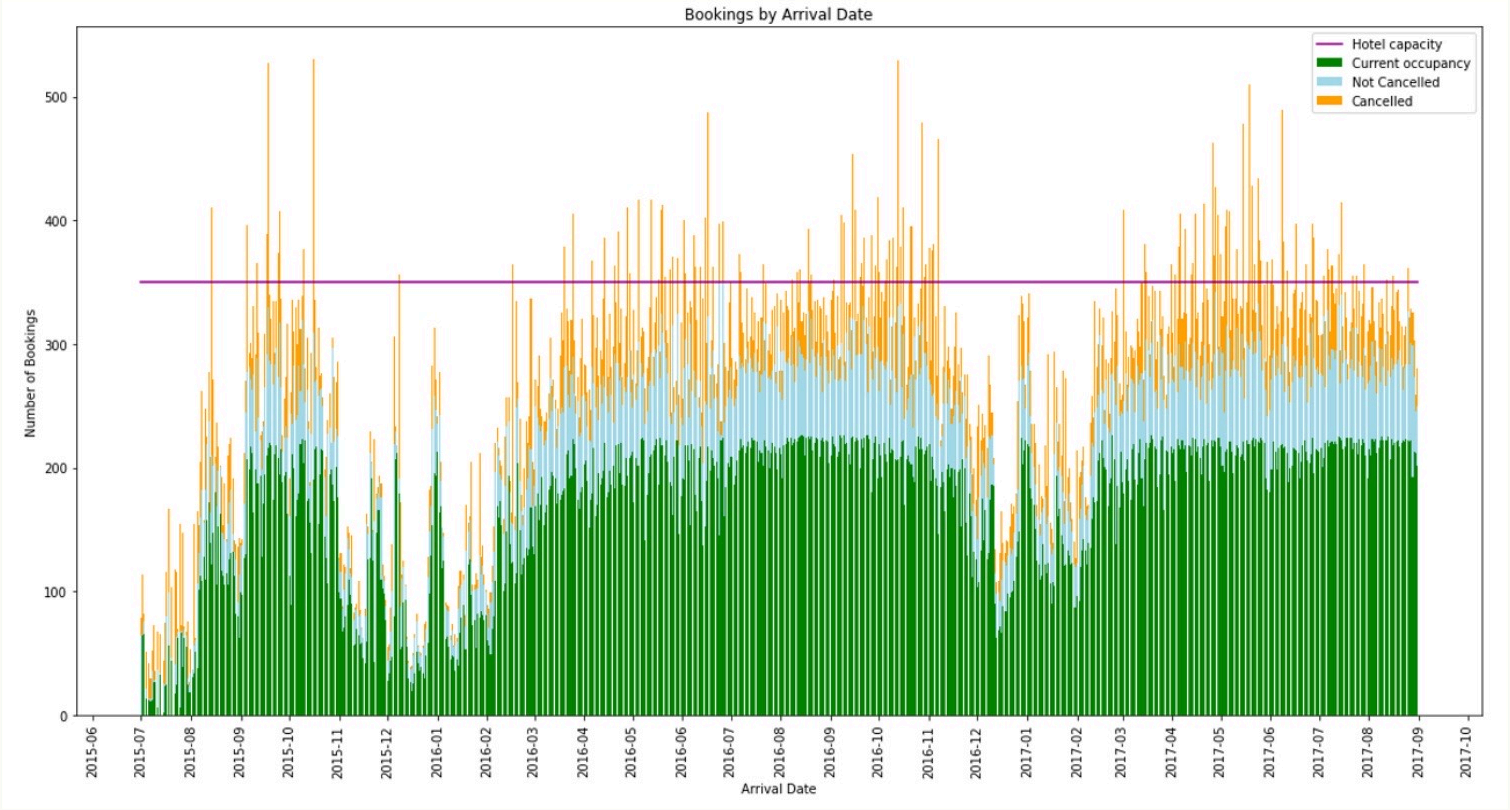

Exploratory Data Analysis
ON CITY HOTEL DATASET

About **City Hotel** Data

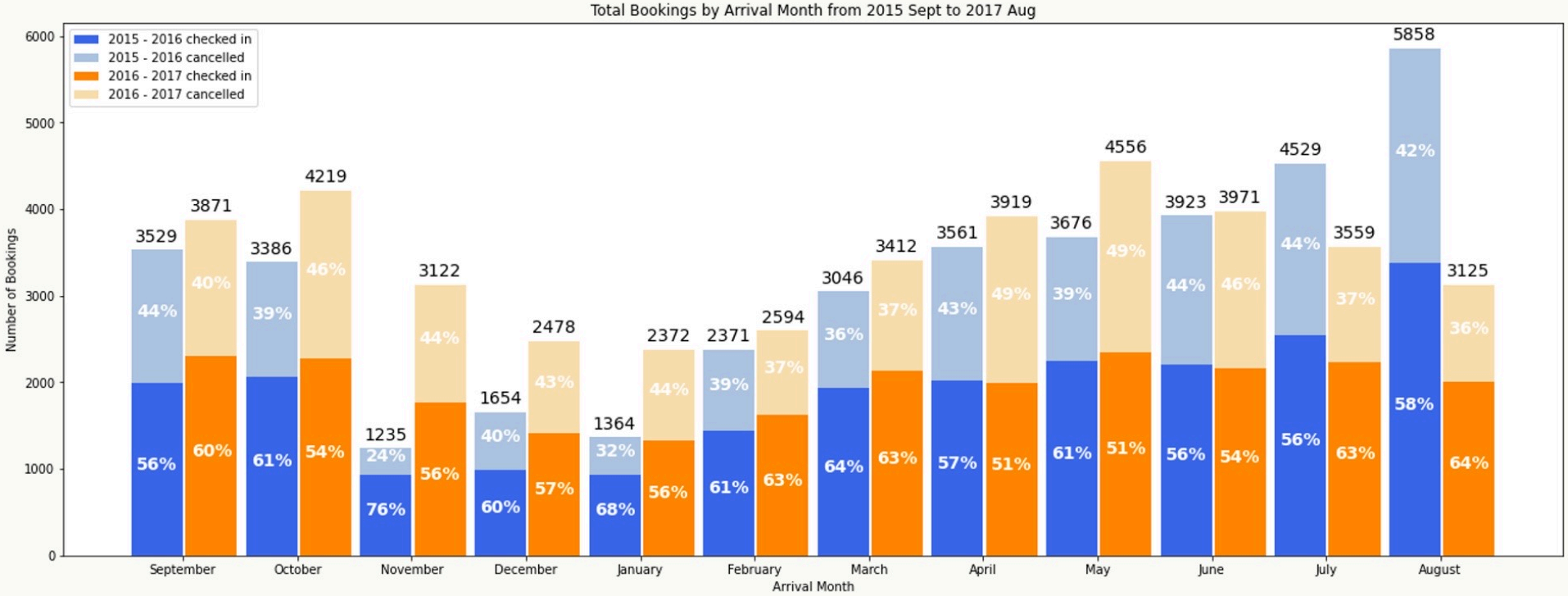
Important Details:

- Room types
- Cancellations
- Guest's arrival date
- Price of room
- Distribution channel

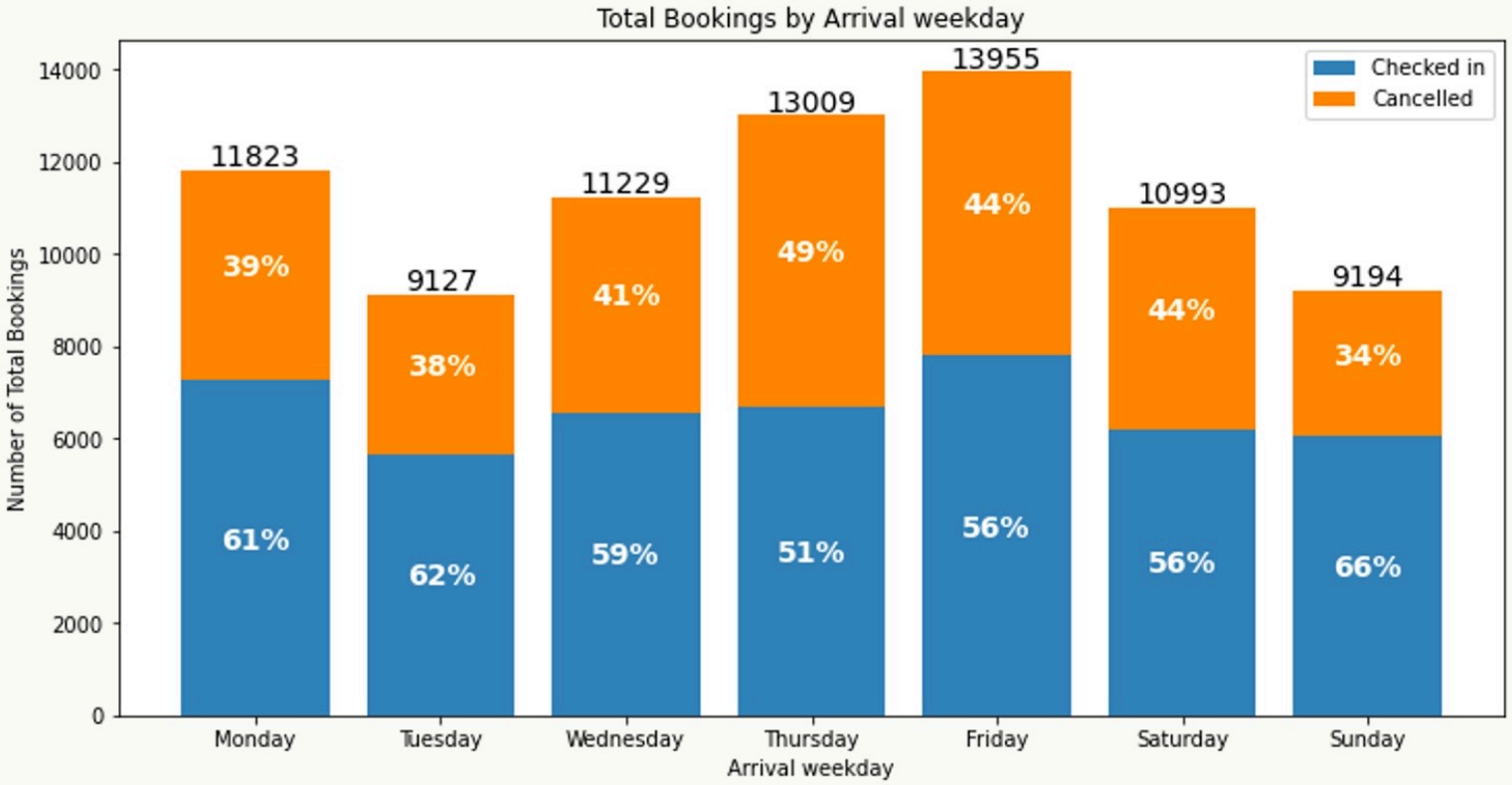
Room Occupancy Trends



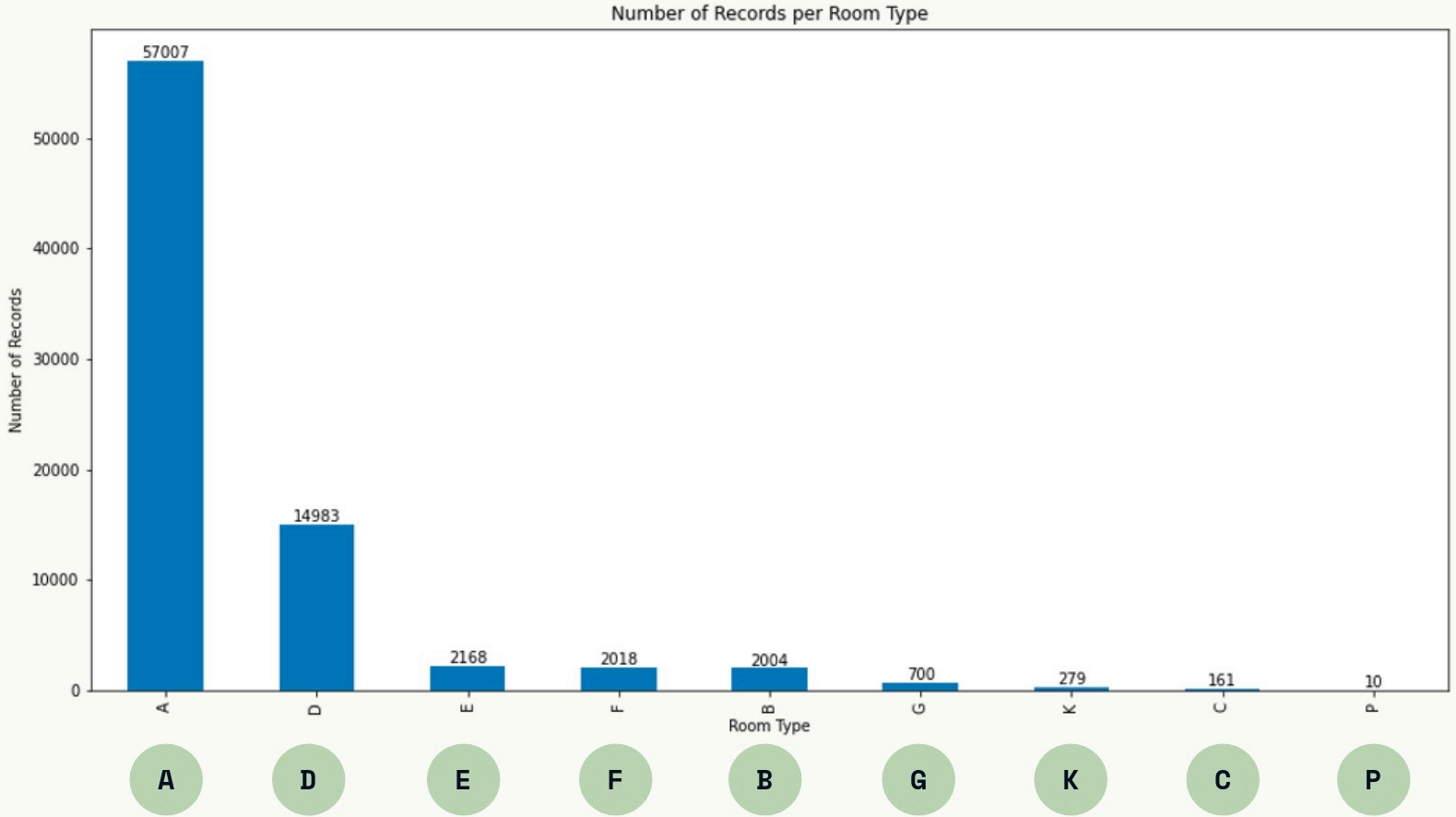
Total Bookings by Month



Bookings vs. Cancellations



Hotel Room Types



Hotel Room Types


Data Preparation

Data Manipulation

To determine:

1. Total capacity of city hotel
2. Bookings vs. Cancellations
3. Peak seasons
4. Premium vs Standard room types

Assuming:

- ✓ Each booked night = 1 booking
- ✗ We will not account for other hotel costs.

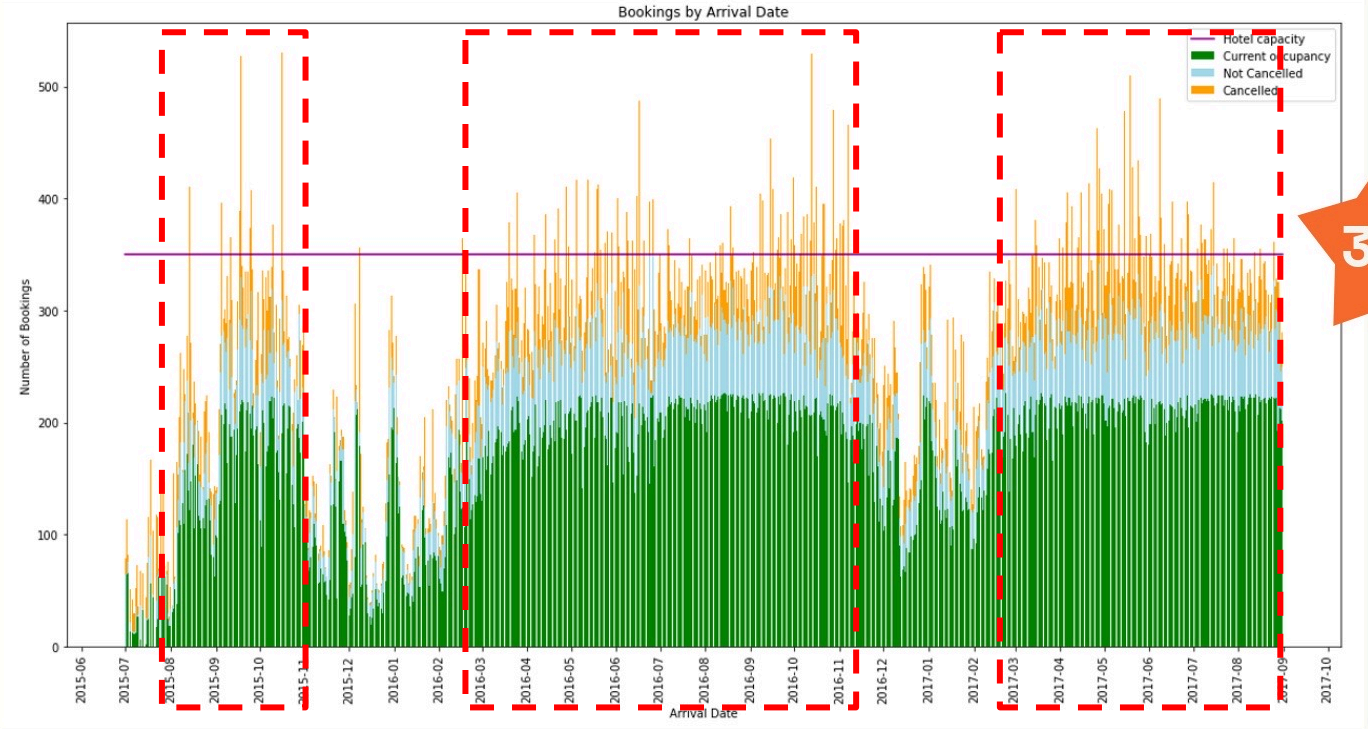
Data Preparation Steps

1. Determine peak periods in a year.
2. Extract records for peak periods in 2015, 2016, 2017.
3. Categorise room types into 2 main types only.
4. Determine total capacity for each room type.
5. Set up data in new dataframe.
6. Apply to Overbooking Model.

Important Variables

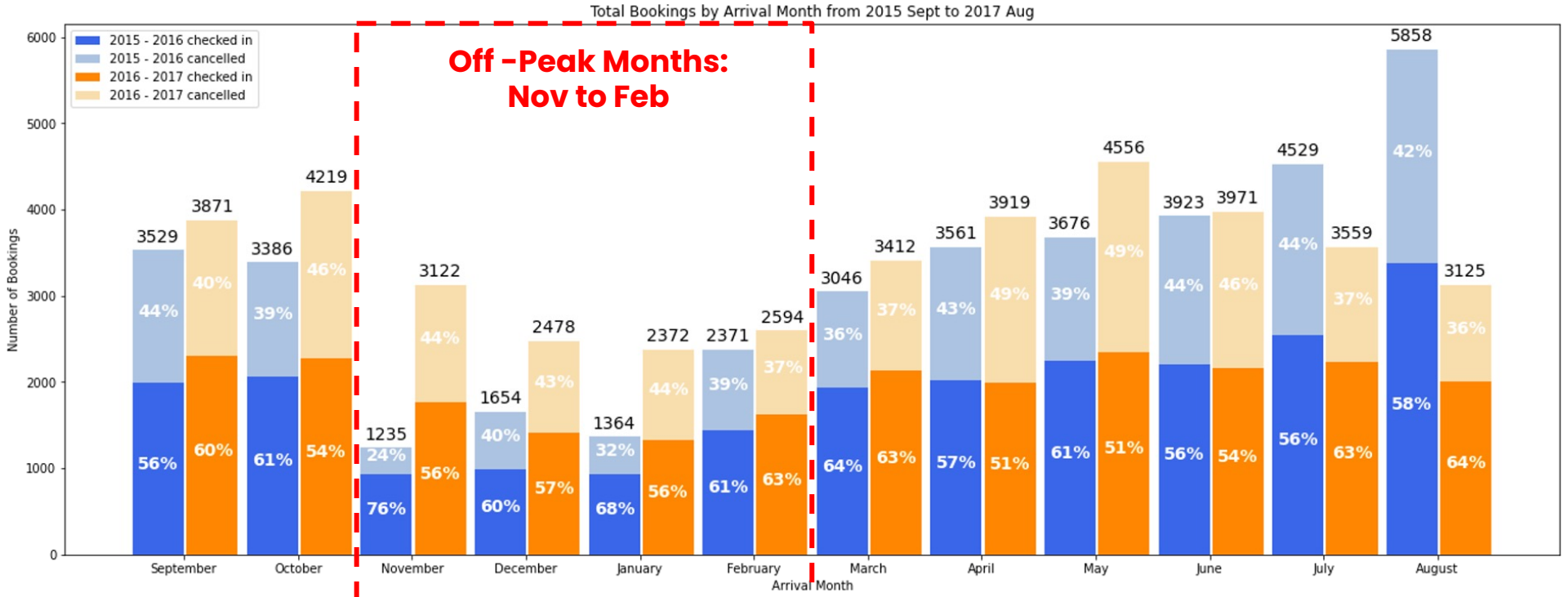
To Find Out:	Variable Name:
Length of Stay	• arrival_date_year • arrival_date_month • arrival_date_day_of_month • stays_in_weekend_nights • stays_in_week_nights
Fulfilment of Booking	• is_cancelled • reservation_status • reserved_room_type • assigned_room_type
Revenue generated	• adr

Step 1 — Determine Peak Periods

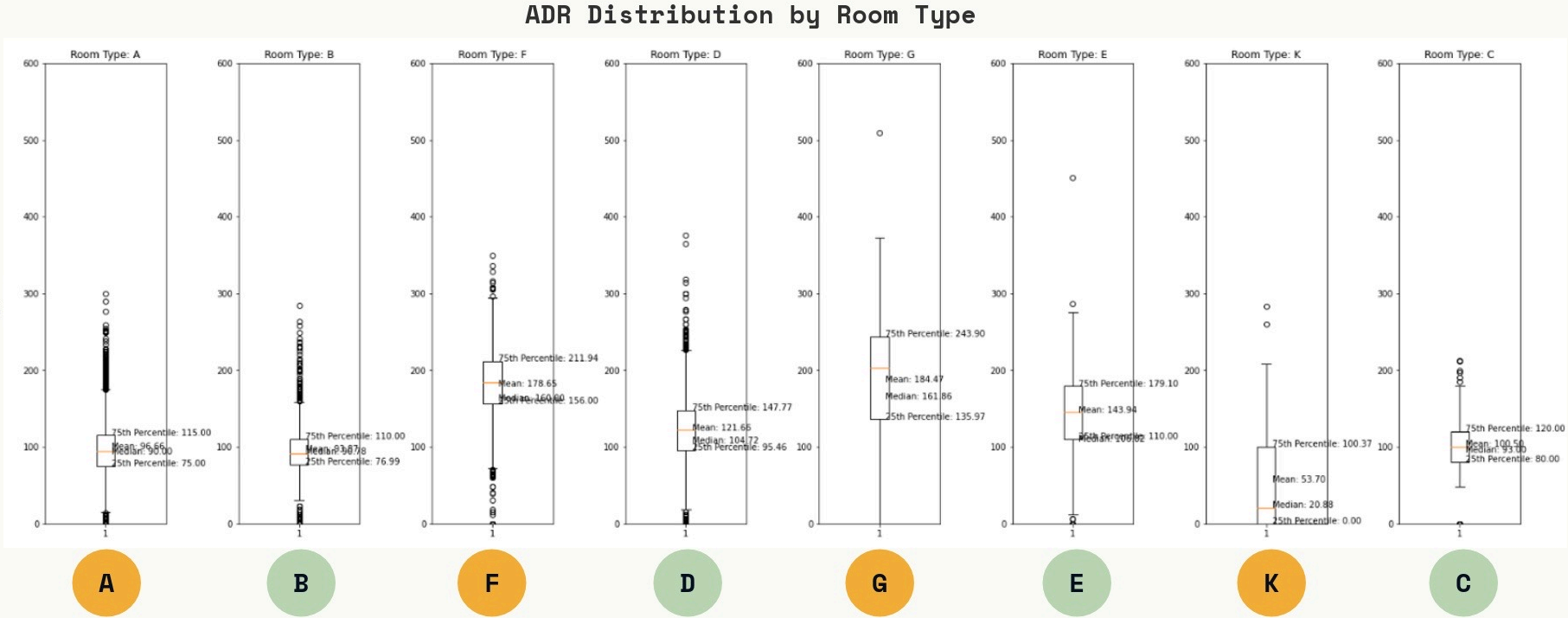


- Total Capacity of Hotel = **350 rooms**

Step 2 — Extract Peak Period Data

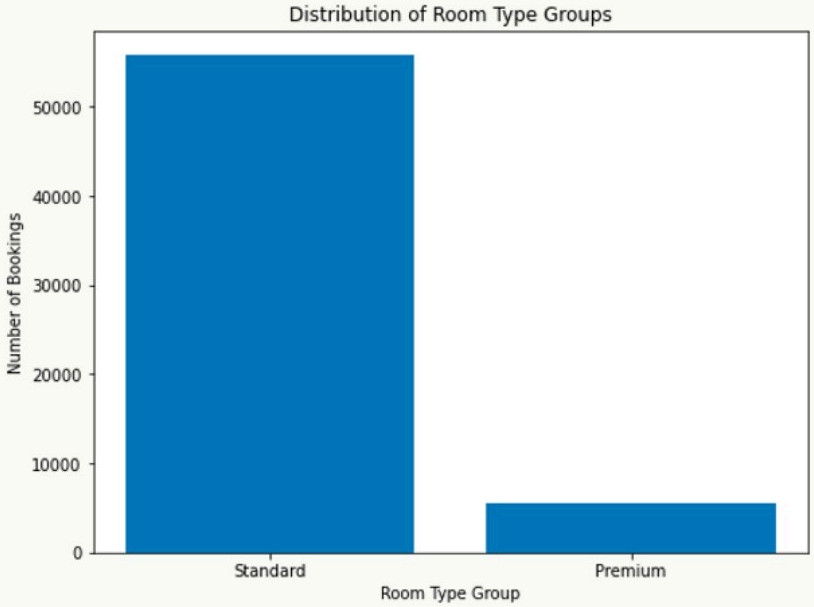


Step 3 — Determine 2 Room Clusters



- Plot of ADR vs. Original Room Types vs. Booking Count

Step 3 — Determine 2 Room Clusters

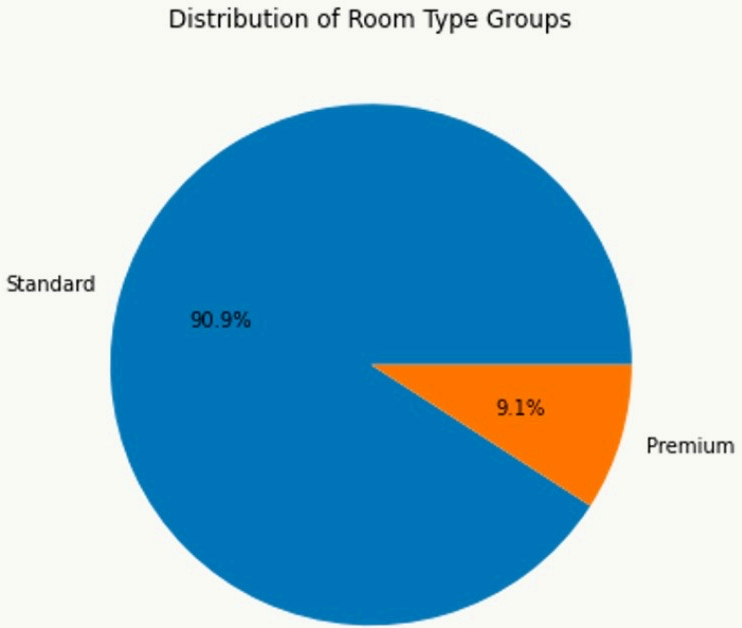


Premium room category — Aim for **10% to 20%** of hotel capacity.

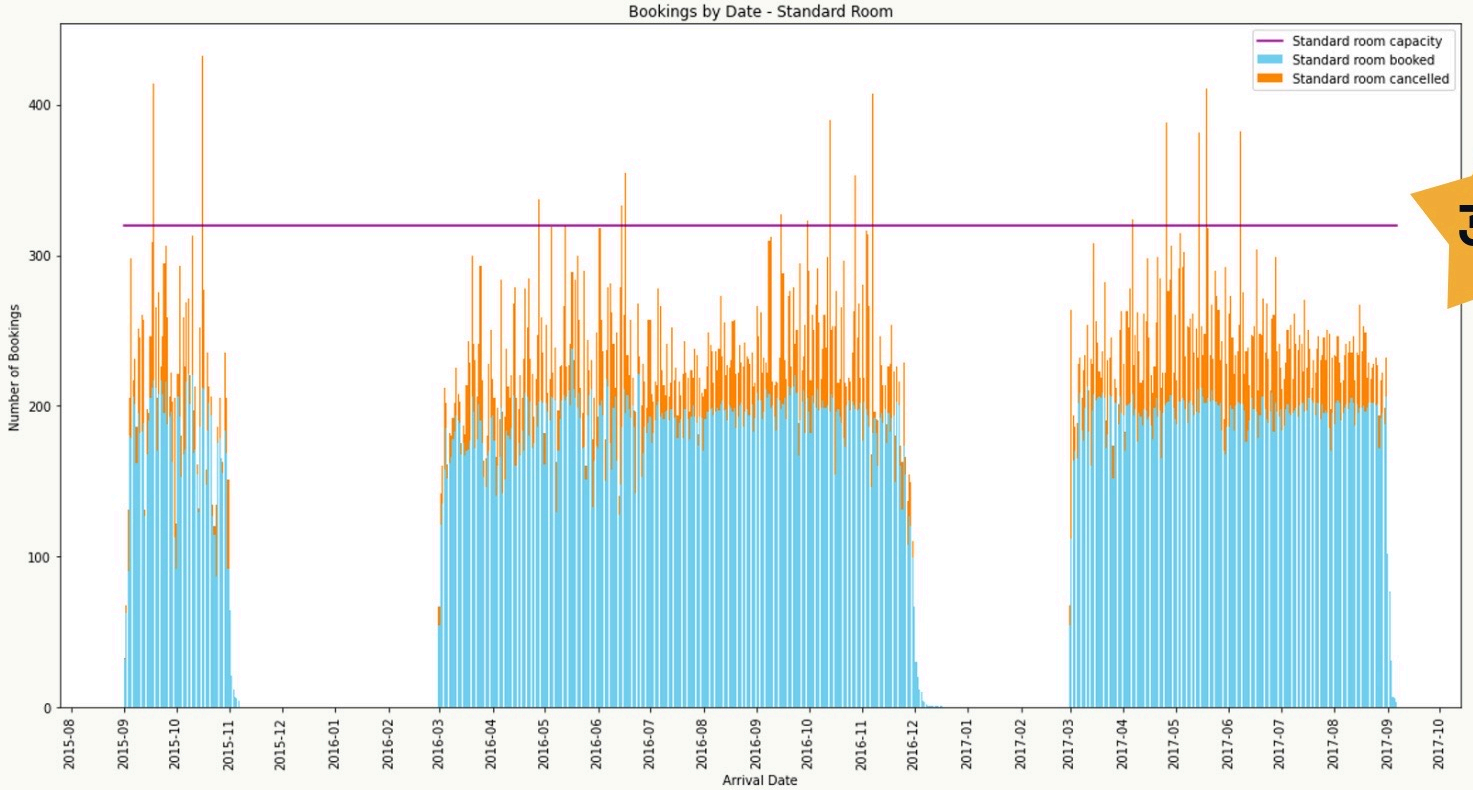
- Plot of Booking Count vs. New Room Categories

Step 3 — Determine 2 Room Clusters

Premium Room Category	Standard Room Category
B	A
E	C
F	D
G	K



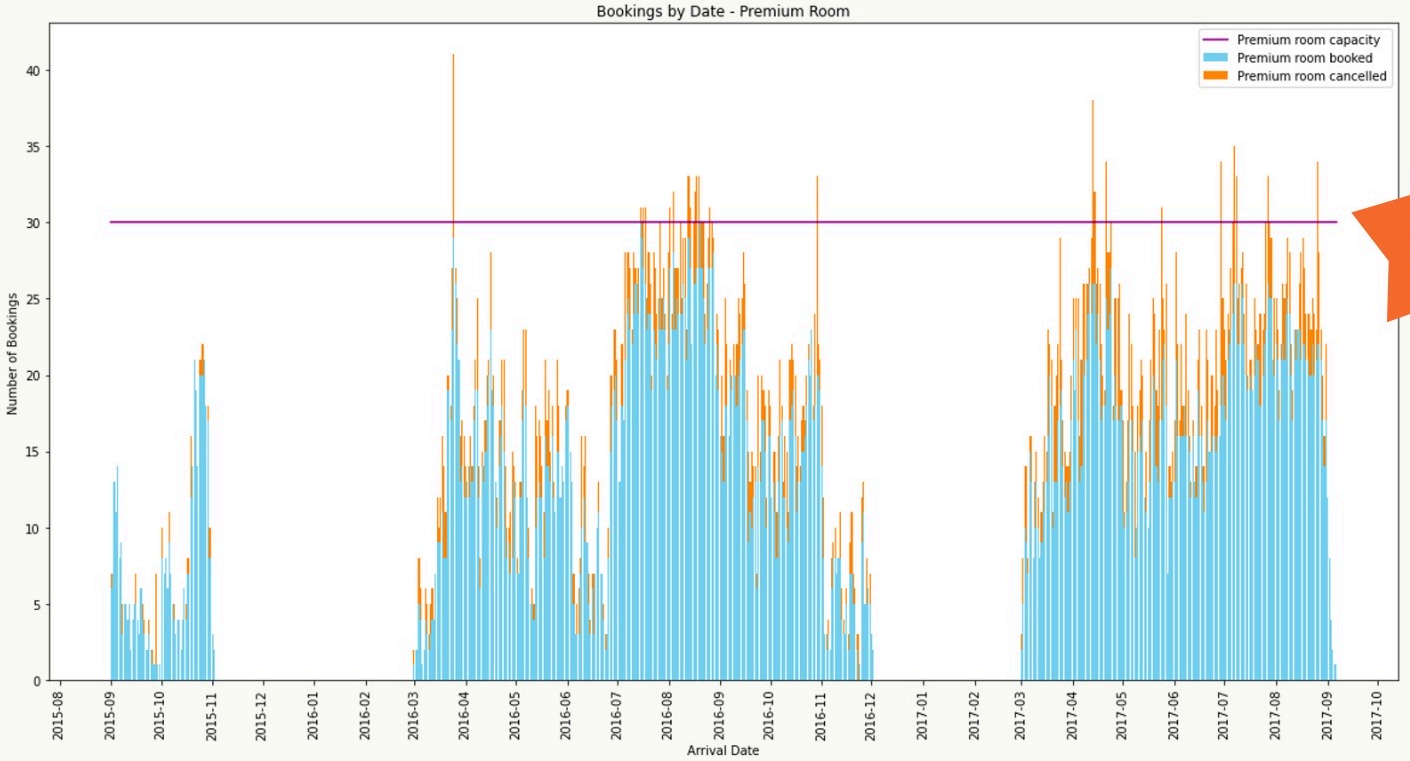
Step 4 — Find Capacity for **Standard** Rooms



320

- Total capacity of **Standard** room type = **320 rooms**

Step 4 — Find Capacity for **Premium** Rooms



30

- Total capacity of **Premium** room type = **30 rooms**

Step 5 — New Dataframe

Day	Capacity (Premium)	Capacity (Standard)	Booked (Premium)	Booked (Standard)	No-shows (Premium)	No-shows (Standard)	Assigned (Premium)	Assigned (Standard)
1								
2								
3								
...								
Total:	N.A.	N.A.	Sum	Sum	Sum	Sum	Sum	Sum

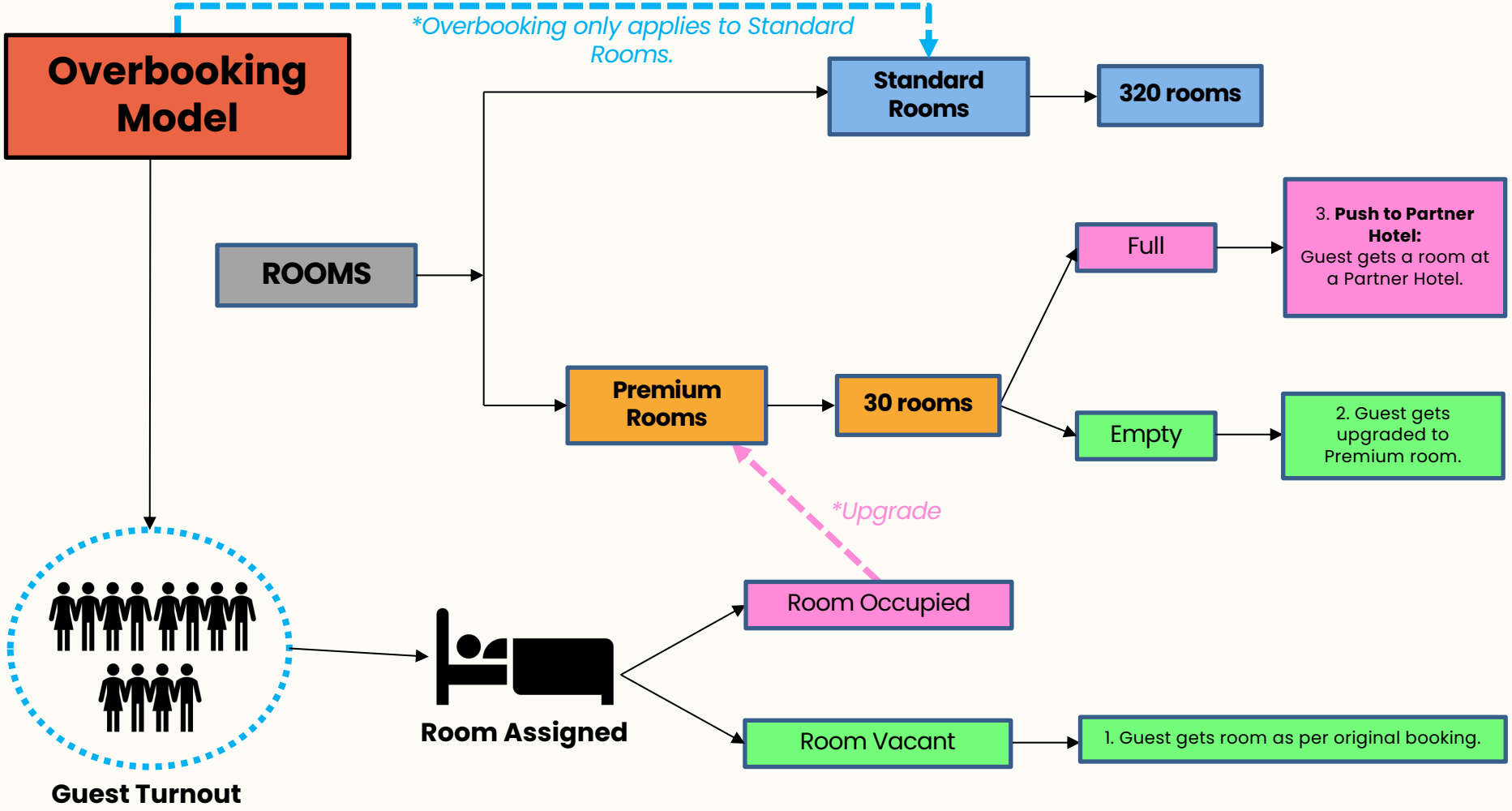
- **Booked** = Total number of bookings per day
- **No-shows** = Total number of cancellations per day
- **Assigned** = Booked – No-shows

Step 5 — New Dataframe

booking_date	premium_capacity	premium_rooms_booked	premium_rooms_canceled	premium_rooms_assigned	standard_capacity	standard_rooms_booked	standard_rooms_canceled	standard_rooms_assigned
2015-09-01	30	7	1	6	320	33	1	32
2015-09-02	30	7	0	7	320	68	5	63
2015-09-03	30	13	0	13	320	131	41	90
2015-09-04	30	11	0	11	320	205	25	180
2015-09-05	30	14	0	14	320	298	119	179
...	...	...	...	...	...	...	...	...
2017-09-02	30	8	0	8	320	77	0	77
2017-09-03	30	4	0	4	320	31	0	31
2017-09-04	30	2	0	2	320	7	0	7
2017-09-05	30	1	0	1	320	6	0	6
2017-09-06	30	1	0	1	320	3	0	3

- Image above shows a partial output of new dataframe.
- Total 549 rows x 8 columns.


Overbooking Model
DURING PEAK SEASON


Overbooking Model
*Overbooking only applies to Standard Rooms.
Standard Rooms
320 rooms
ROOMS
Premium Rooms
30 rooms
Full
3. Push to Partner Hotel: Guest gets a room at a Partner Hotel.
Empty
2. Guest gets upgraded to Premium room.
*Upgrade
Guest Turnout
Room Assigned
Room Occupied
Room Vacant
1. Guest gets room as per original booking.

What do we earn from Overbooking?

- For No-shows, the hotel **does not give refunds**.
- The hotel keeps all charges that have been paid by customer.

Building the Overbooking Model

To determine:

1. Margin for overbooking
2. ADR value for Standard Rooms
3. Risk penalty amount for Standard room type

Assumptions:

- Excluding early departures
- Excluding overstays
- No refunds

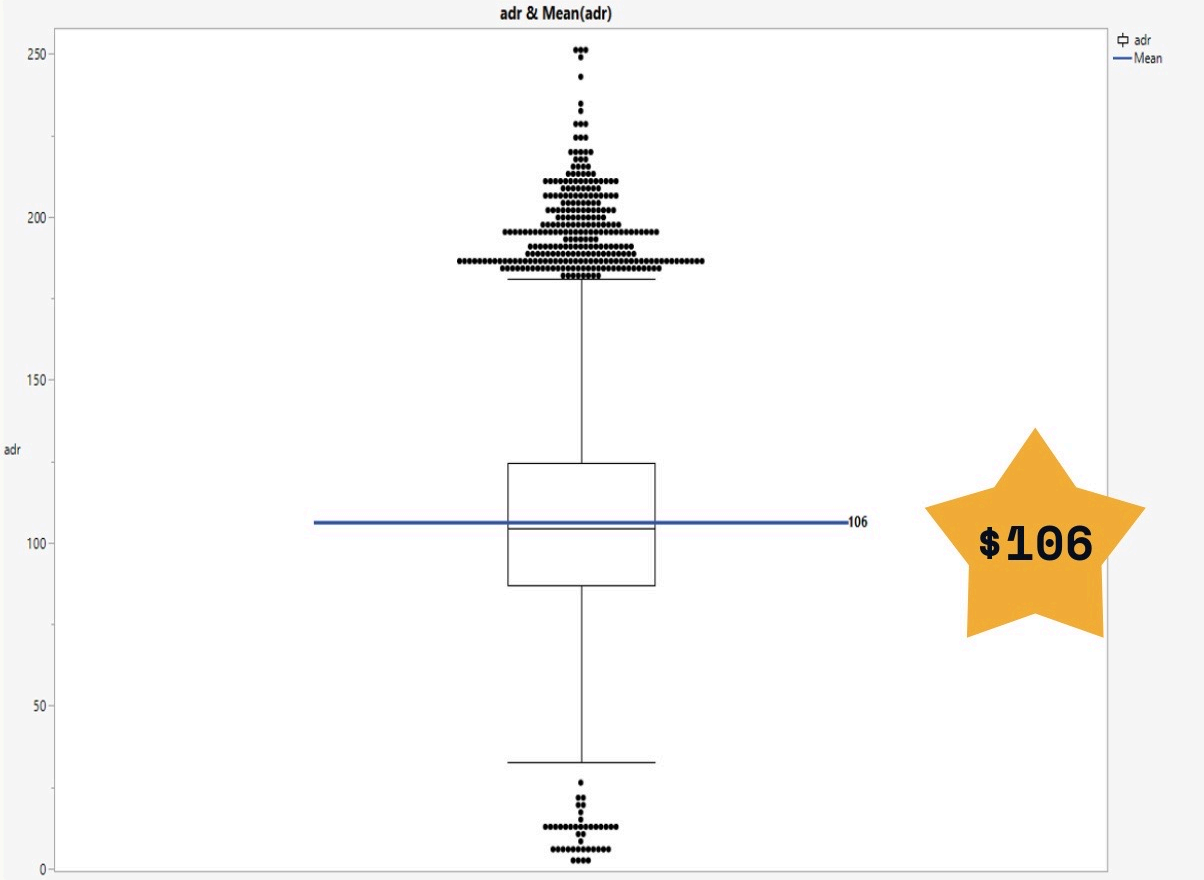
Margin for Overbooking

$$\text{Customer Service Level (CSL)} = \frac{\textit{No. of days where realised demand in peak season} \leq \textit{Capacity}}{\textit{Total number of days in peak season}} \times 100\%$$

For example, to maintain a **CSL of 80%**:

- If there are **275 peak days** in a year, we can only truly exceed the hotel's capacity on a maximum of **55 days**.

ADR per Standard Room



Refresh your Memory!

Day	Capacity (Premium)	Capacity (Standard)	Booked (Premium)	Booked (Standard)	No-shows (Premium)	No-shows (Standard)	Assigned (Premium)	Assigned (Standard)
1								
2								
3								
...								
Total:	N.A.	N.A.	Sum	Sum	Sum	Sum	Sum	Sum

- **Booked** = Total number of bookings per day
- **No-shows** = Total number of cancellations per day
- **Assigned** = Booked – No-shows

Building the Overbooking Model

- Each booking is considered a **Bernoulli Problem** → Show-up or No-show.

$$\sigma = \sqrt{\frac{pq}{X}}$$

σ = *Standard deviation of the proportion of No-shows*
p = *Probability of a No-show*
q = *Probability of a room assigned*
X = *Number of room bookings*

Building the Overbooking Model

For Standard room type:

$$\textit{Probability of No-show, } p_{Standard} = \frac{Sum\ of\ No{-}show\ (Standard)}{Sum\ of\ Booked\ (Standard)} = 0.1972$$

$$\textit{Probability of Standard room being assigned, } q_{Standard} = 1 - p_{Standard} = 0.8028$$

Cancellation rate for standard rooms

In [77]:
```
standard_cancellation = df_city_total_sum['standard_rooms_canceled'] / df_city_total_sum['standard_rooms_booked']
standard_cancellation = round(standard_cancellation, 4)
p_standard = standard_cancellation
q_standard = round(1 - p_standard,4)
p_standard, q_standard
```

Out[77]: (0.1972, 0.8028)

Building the Overbooking Model

- Since the sample size is sufficiently large, we can conclude that the proportion of no-shows is **Normal**, as per *Central Limit Theorem*.

By substituting the terms in $\sigma = \sqrt{\frac{pq}{X}}$,
we can derive a new equation with only 1 unknown, *X*:

$$X\left[1-\left(p-z\sqrt{\frac{pq}{X}}\right)\right]=C$$

Where *p, q, z* and *C* are known constants.

p = *Probability of a No-show*
q = *Probability of a room assigned*
z = *Z-score value corresponding to Customer Service Level*
C = *Maximum number of Standard rooms in hotel*

X = Number of room bookings allowed

Building the Overbooking Model

- To solve for the ***unknown X***, we apply further mathematical transformations to get a quadratic equation:

$$X[(1-p)^2 X^2 - [2C(1-p) + z^2pq]X + C^2] = 0$$

p = *Probability of a No-show = 0.1972*
q = *Probability of a room assigned = 0.8028*
z = *Z-score value corresponding to Customer Service Level*
C = *Maximum number of Standard rooms in hotel = 320 rooms*

X = Number of room bookings allowed

Different Customer Service Levels

- Referencing to a ***z-score*** **percentile table** for normal distribution:

Percentile	z-Score	Percentile	z-Score	Percentile	z-Score
13	-1.126	46	-0.1	79	0.806
14	-1.08	47	-0.075	80	0.842
15	-1.036	48	-0.05	81	0.878
16	-0.994	49	-0.025	82	0.915
17	-0.954	50	0	83	0.954
18	-0.915	51	0.025	84	0.994
19	-0.878	52	0.05	85	1.036
20	-0.842	53	0.075	86	1.08
21	-0.806	54	0.1	87	1.126
22	-0.772	55	0.126	88	1.175
23	-0.739	56	0.151	89	1.227
24	-0.706	57	0.176	90	1.282
25	-0.674	58	0.202	91	1.341
26	-0.643	59	0.228	92	1.405
27	-0.613	60	0.253	93	1.476
28	-0.583	61	0.279	94	1.555
29	-0.553	62	0.305	95	1.645
30	-0.524	63	0.332	96	1.751
31	-0.496	64	0.358	97	1.881
32	-0.468	65	0.385	98	2.054
33	-0.44	66	0.412	99	2.326

Table taken from Pindling.org

Different Customer Service Levels

CSL	z-score
80%	0.842
85%	1.036
90%	1.282
95%	1.645
98%	2.054

Building the Overbooking Model

- Now, we can solve for X through a general quadratic equation.

At 80% service level, using $z = 0.842$,

$$a = 0.644$$
$$b = -512.11$$
$$c = 102{,}400$$

$$X_{\_80} = \left(\frac{-b-\sqrt{b^2-4ac}}{2a}\right) = 390.36$$

Case 1: Service level = 80%, z = 0.842

In [94]:
```
# parameters = z_80, p_standard, q_standard, C = 320 (standard room cap), X = max allowed booking
z_80 = 0.842

# using quadratic eqn: X[(1-p)^2 X^2 - [2C(1-p) + (z^2)*pq]X + C^2] = 0
# X = 0 is one of the roots

a = (1-p_standard) ** 2
b = - (2 * 320 * (1-p_standard) + (z_80 ** 2) * p_standard * q_standard)
c = 320 ** 2
```

In [95]:
```
x_80 = (-b - (b**2 - 4 * a * c)**0.5)/(2*a)
x_80
```

Out[95]: 390.3598116255052

Building the Overbooking Model

CSL	Number of bookings allowed, X
80%	390
85%	388
90%	386
95%	382
98%	379

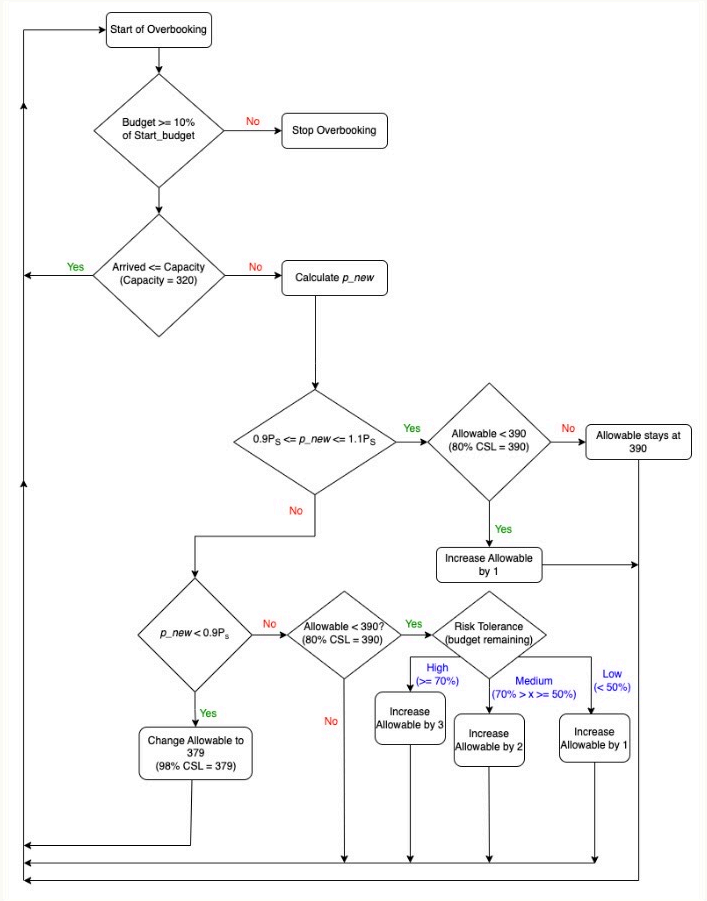
Penalty for Overbooking

Type of Penalty	Cost ($)
Re-accomodate guest at Partner Hotel	$200/room
Reputational Damage	$100/booking

- The penalty will be taken out of the allocated **starting budget**.

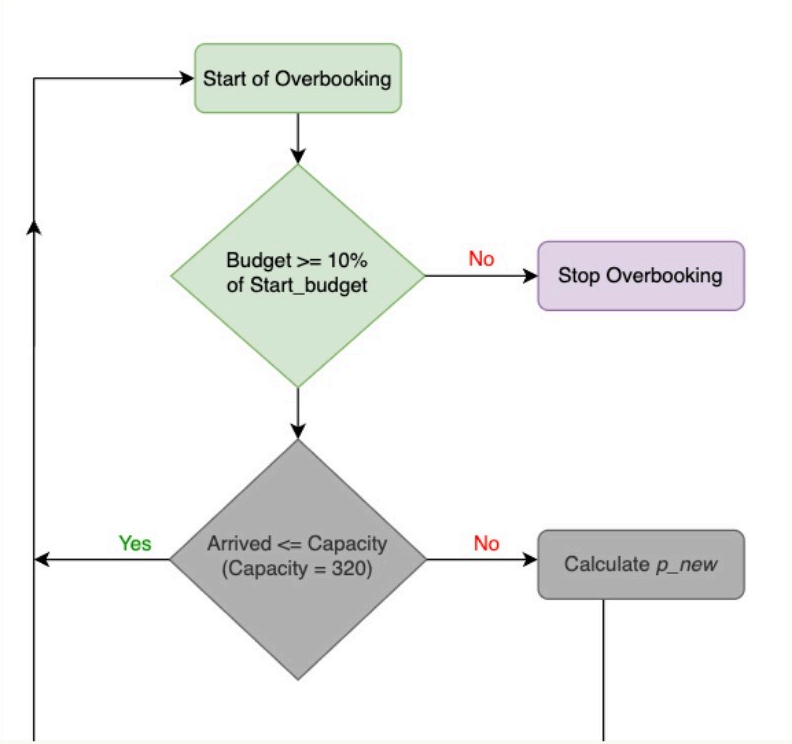

Evaluation

Flow Chart for Overbooking Strategy



Closing the loop — Scenario 1

- What happens as the amount of budget leftover changes?

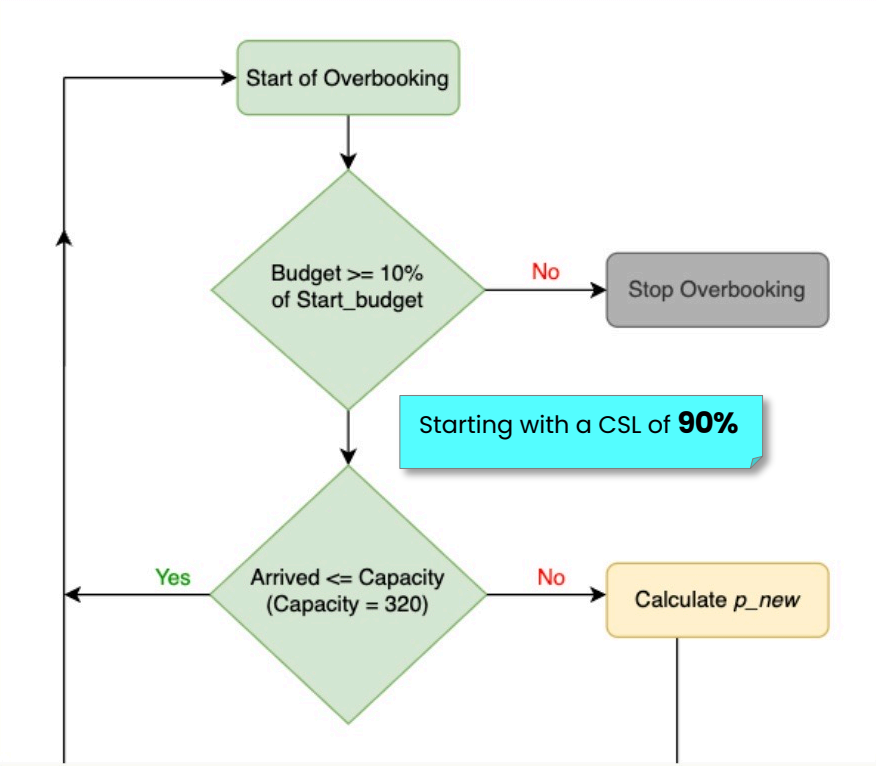


E.g.
There is **$10,000** left out of the $150,000 starting budget.

Why?
Adjusting an overbooking model may lead to reduction in the remaining budget, because potential compensations for overbooked guests may surpass hotel capacity in the future.

Closing the loop — Scenario 2

- What happens when number of **actual** guests arrived <= hotel capacity?



YES scenario:
- Arrived = 300
- Capacity = 320

NO scenario:
- Booked = 386
- Arrived = 325
- Capacity = 320
- Recommended Allowable at 90% CSL = 386

Refresh your Memory!

For Standard room type:

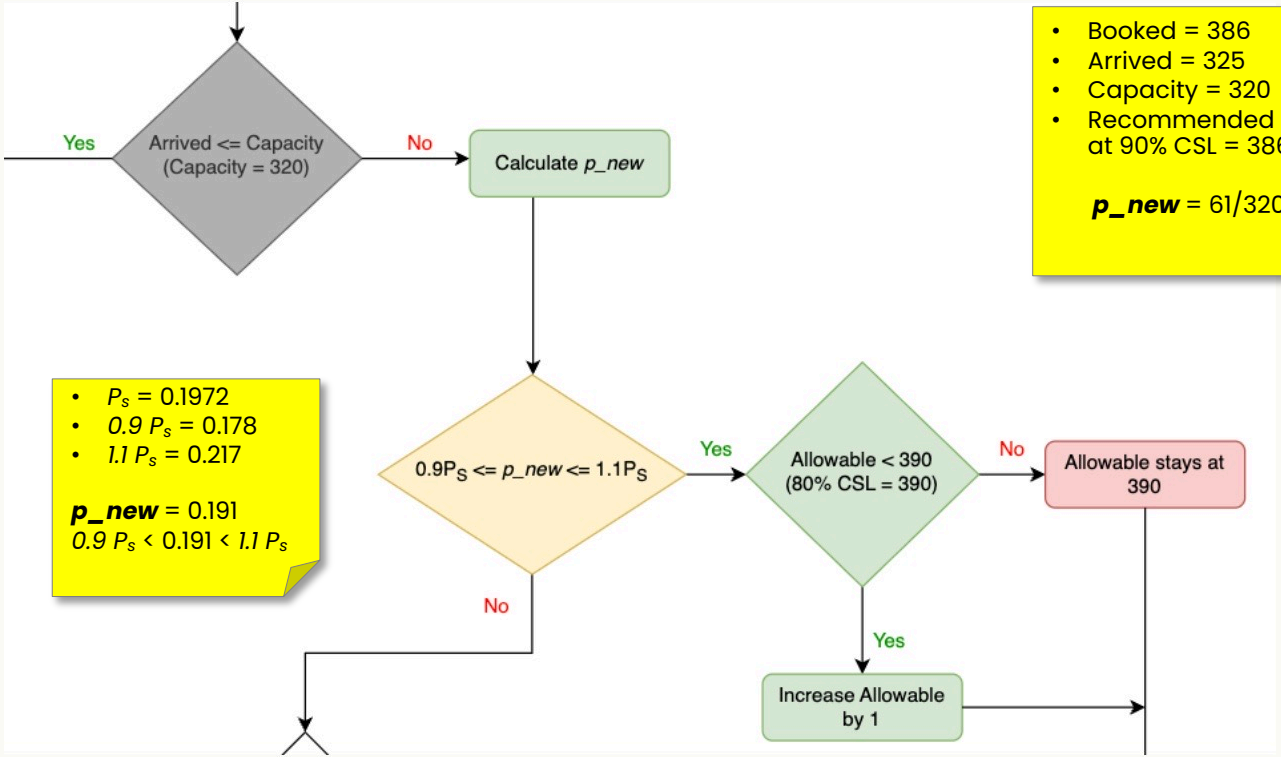
Probability of No-show, $p_{Standard} = \frac{Sum\ of\ No-show\ (Standard)}{Sum\ of\ Booked\ (Standard)} = 0.1972$

Probability of Standard room being assigned, $q_{Standard} = 1 - p_{Standard} = 0.8028$

- *Note – '*p*' is the probability of **No-show**!

Closing the loop — Scenario 3

- What happens if *p_new* **falls within** the safe range?



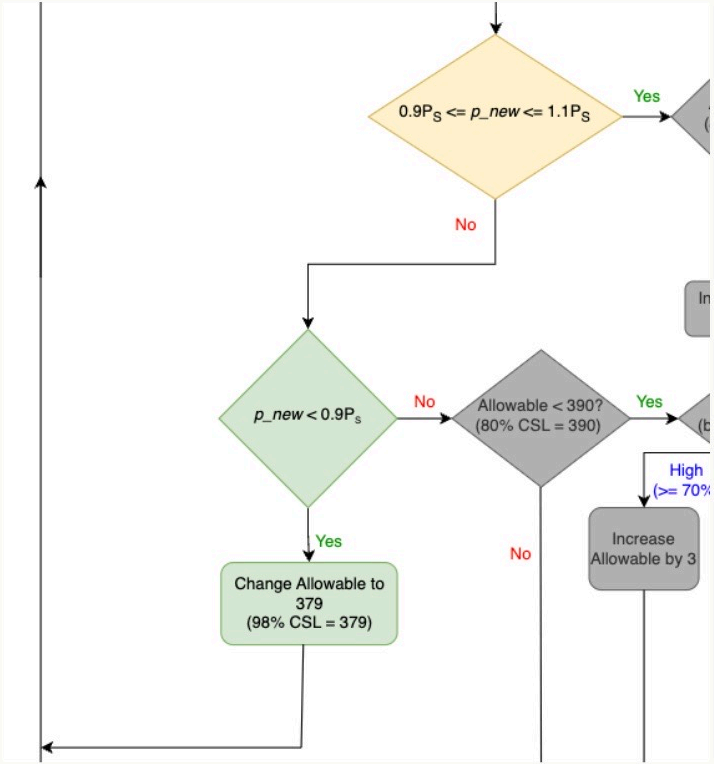
Closing the loop — Scenario 4

- What happens if *p_new* is **< 90%** of *p_standard*? (meaning probability of no-shows is lower)

- P_s = 0.1972
- *0.9* P_s = 0.178
- *1.1* P_s = 0.217

p_new = 0.113
0.113 < 0.9 P_s

p_new is now lower.



- Booked = 386
- Arrived = 350
- Capacity = 320
- Recommended Allowable at 90% CSL = 386

p_new = 36/320 = 0.113

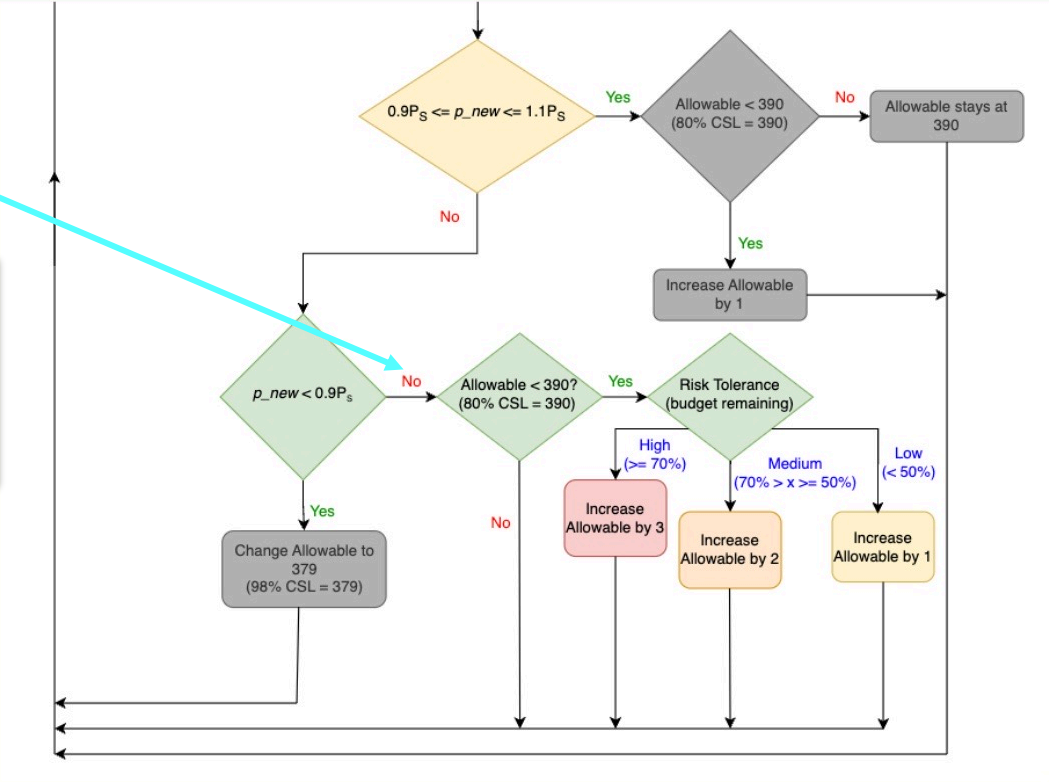
Closing the loop — Scenario 5

- What happens if *p_new* is **> 110%** of *p_standard*? (meaning probability of no-shows is higher)

Meaning *p_new* **> 1.1 P_s**

- Booked = 400
- Arrived = 325
- Capacity = 320
- Recommended Allowable at 90% CSL = 386

p_new = 75/320 = 0.234



- P_s = 0.1972
- *0.9* P_s = 0.178
- *1.1* P_s = 0.217

p_new = 0.234
0.234 > *1.1* P_s

p_new is now higher.

Visualisation for Revenue

Parameters		
Max Capacity (Standard Rooms)	320	
Peak days in a year (1 season)	275	
ADR per standard room	$106	
Total costs (TC)	$300	
TC_partner hotel	$200	
TC_reputational damage	$100	
Starting budget	$150,000	
Customer Service Level = 98 %	379	*max rooms to overbook*
Customer Service Level = 90 %	386	*max rooms to overbook*
Customer Service Level = 80 %	390	*max rooms to overbook*

Case 1 *Customer service - 98%* ***Arrived <= Capacity***		**(A)** **Model - Overbooking Limit**	**(B)** **Rooms Booked**	**(C)** **Actual Arrived**	**(D)** **Actual Revenue** *(B) * ADR*	**(E)** **Optimal Revenue** *Max Capacity * ADR*	**(F)** **Total Costs** *TC * Exceeded Capacity*	**(G)** **Actual Profits** *(D) - (F)*	**(H)** **Cancellation Rate today** *((B)-(C)) / (B)*	
WITH OB MODEL	Day 1	386	386	315	$40,916	$33,920	0	$40,916	18%	CLOSE LOOP + UPDATE MODEL FOR DAY 2
WITHOUT OB MODEL	Day 1	320	320	315	$33,920	$33,920	0	$33,920	2%	
WITH OB MODEL	Day 2									
WITHOUT OB MODEL	Day 2									

Case 2 *Customer service - 98%* ***Arrived > Capacity***		**(A)** **Model - Overbooking Limit**	**(B)** **Rooms Booked**	**(C)** **Actual Arrived**	**(D)** **Actual Revenue** *(B) * ADR*	**(E)** **Optimal Revenue** *Max Capacity * ADR*	**(F)** **Total Costs** *TC * Exceeded Capacity*	**(G)** **Actual Profits** *(D) - (F)*	**(H)** **Cancellation Rate today** *((B)-(C)) / (B)*	
WITH OB MODEL	Day 1	386	386	350	$40,916	$33,920	9000	$31,916	9%	CLOSE LOOP + UPDATE MODEL FOR DAY 2
WITHOUT OB MODEL	Day 1	320	320	320	$33,920	$33,920	0	$33,920	0%	
WITH OB MODEL	Day 2									
WITHOUT OB MODEL	Day 2									

Visualisation for Revenue

- What happens when number of **actual** guests arrived <= hotel capacity?

Parameters		
Max Capacity (Standard Rooms)	320	
Peak days in a year (1 season)	275	
ADR per standard room	$106	
Total costs (TC)	$300	
TC_partner hotel	$200	
TC_reputational damage	$100	
Starting budget	$150,000	
Customer Service Level = 98 %	379	*max rooms to overbook*
Customer Service Level = 90 %	386	*max rooms to overbook*
Customer Service Level = 80 %	390	*max rooms to overbook*

Case 1 *Customer service - 98%* *Arrived <= Capacity*		**(A)** Model - Overbooking Limit	**(B)** Rooms Booked	**(C)** Actual Arrived	**(D)** Actual Revenue *(B) * ADR*	**(E)** Optimal Revenue *Max Capacity * ADR*	**(F)** Total Costs *TC * Exceeded Capacity*	**(G)** Actual Profits *(D) - (F)*	**(H)** Cancellation Rate today *((B)-(C)) / (B)*	
WITH OB MODEL	Day 1	386	386	315	$40,916	$33,920	0	$40,916	18%	CLOSE LOOP + UPDATE MODEL FOR DAY 2
WITHOUT OB MODEL	Day 1	320	320	315	$33,920	$33,920	0	$33,920	2%	
WITH OB MODEL	Day 2									
WITHOUT OB MODEL	Day 2									

OB allows extra earnings of $6996.

Visualisation for Revenue

- What happens when number of **actual** guests arrived **>** hotel capacity?

Parameters		
Max Capacity (Standard Rooms)	320	
Peak days in a year (1 season)	275	
ADR per standard room	$106	
Total costs (TC)	$300	
TC_partner hotel	$200	
TC_reputational damage	$100	
Starting budget	$150,000	
Customer Service Level = 98 %	379	*max rooms to overbook*
Customer Service Level = 90 %	386	*max rooms to overbook*
Customer Service Level = 80 %	390	*max rooms to overbook*

Case 2 *Customer service - 98%* *Arrived > Capacity*		(A) Model - Overbooking Limit	(B) Rooms Booked	(C) Actual Arrived	(D) Actual Revenue *(B) * ADR*	(E) Optimal Revenue *Max Capacity * ADR*	(F) Total Costs *TC * Exceeded Capacity*	(G) Actual Profits *(D) - (F)*	(H) Cancellation Rate today *((B)-(C)) / (B)*	
WITH OB MODEL	Day 1	386	386	386	$40,916	$33,920	19800	$21,116	0%	CLOSE LOOP + UPDATE MODEL
WITHOUT OB MODEL	Day 1	320	320	320	$33,920	$33,920	0	$33,920	0%	FOR DAY 2
WITH OB MODEL	Day 2									
WITHOUT OB MODEL	Day 2									

At this point, OB still generates revenue, but lesser.

Visualisation for Allowable OB Days

- As time passes, how do we decide if we can continue overbooking ?

Customer Service Level (C.S.L)	90.00%	90.15%	90.48%	90.81%	90.77%	90.74%	90.71%	90.67%	90.64%	90.98%	91.32%	91.67%	92.02%	91.98%	91.95%	92.31%	92.66%
	Starting point	**EOD 1**	**EOD 2**	**EOD 3**	**EOD 4**	**EOD 5**	**EOD 6**	**EOD 7**	**EOD 8**	**EOD 9**	**EOD 10**	**EOD 11**	**EOD 12**	**EOD 13**	**EOD 14**	**EOD 15**	**EOD 16**
Is Arrived > Capacity (320)?	NIL	0	1	1	0	0	0	0	0	1	1	1	1	0	0	1	1
Allowable days where Arrived > Capacity	27	27	26	25	25	25	25	25	25	24	23	22	21	21	21	20	19
Days where Arrived <= Capacity	248	247	247	247	246	245	244	243	242	242	242	242	242	241	240	240	240
Remaining no. of days in Peak season	275	274	273	272	271	270	269	268	267	266	265	264	263	262	261	260	259

0	Arrived < Capacity (320)
1	Arrived > Capacity (320)

- If **CSL increases** = Hotel has to **reduce** overbooking.
- If **CSL decreases** = Hotel can **increase** overbooking.

Visualisation for Allowable OB Days

- If Arrived is consistently <= Capacity of hotel:

Customer Service Level (C.S.L)	90.00%	90.15%	90.11%	90.07%	90.04%	90.00%	89.96%	89.93%	89.89%	89.85%	89.81%	89.77%	89.73%	89.69%	89.66%	89.62%	89.58%

	Starting point	EOD 1	EOD 2	EOD 3	EOD 4	EOD 5	EOD 6	EOD 7	EOD 8	EOD 9	EOD 10	EOD 11	EOD 12	EOD 13	EOD 14	EOD 15	EOD 16
Is Arrived > Capacity (320)?	NIL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Allowable days where Arrived > Capacity	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27	27
Days where Arrived <= Capacity	248	247	246	245	244	243	242	241	240	239	238	237	236	235	234	233	232
Remaining no. of days in Peak season	275	274	273	272	271	270	269	268	267	266	265	264	263	262	261	260	259

0	Arrived < Capacity (320)
1	Arrived > Capacity (320)

- **CSL decreases** steadily = Hotel can **increase** overbooking.

Visualisation for Allowable OB Days

- If Actual Arrived is consistently **>** Capacity of hotel:

Customer Service Level (C.S.L)	90.00%	90.51%	90.84%	91.18%	91.51%	91.85%	92.19%	92.54%	92.88%	93.23%	93.58%	93.94%	94.30%	94.66%	95.02%	95.38%	95.75%
	Starting point	**EOD 1**	**EOD 2**	**EOD 3**	**EOD 4**	**EOD 5**	**EOD 6**	**EOD 7**	**EOD 8**	**EOD 9**	**EOD 10**	**EOD 11**	**EOD 12**	**EOD 13**	**EOD 14**	**EOD 15**	**EOD 16**
Is Arrived > Capacity (320)?	NIL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Allowable days where Arrived > Capacity	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11
Days where Arrived <= Capacity	248	248	248	248	248	248	248	248	248	248	248	248	248	248	248	248	248
Remaining no. of days in Peak season	275	274	273	272	271	270	269	268	267	266	265	264	263	262	261	260	259

0	Arrived < Capacity (320)
1	Arrived > Capacity (320)

- **CSL increases** steadily = Hotel has to **reduce** overbook.

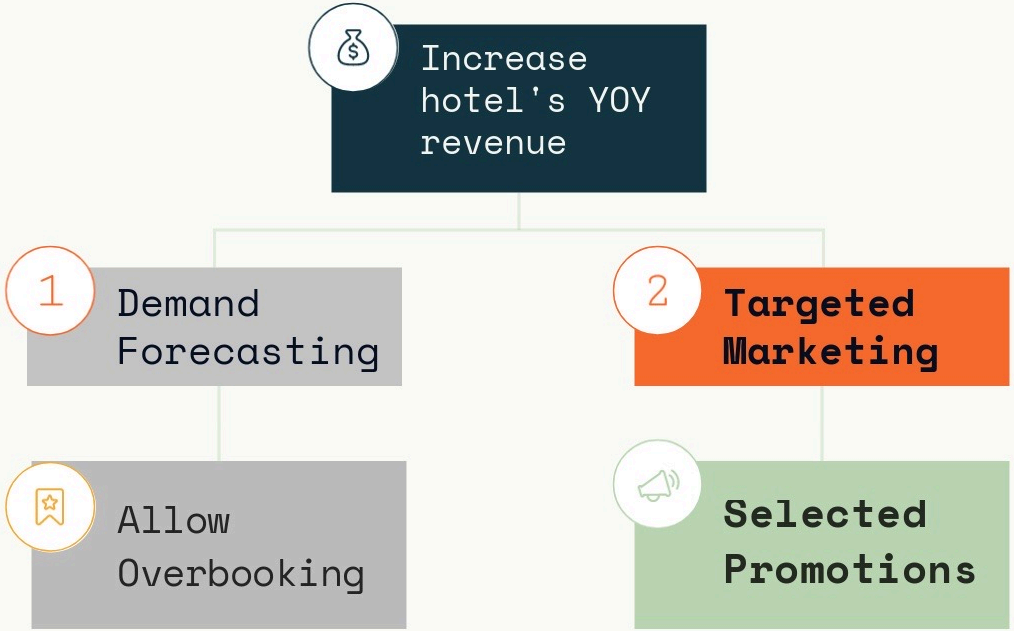

Future
Considerations

Pain Points in Hotel Overbooking

- Hotel overbooking models are **more complicated**.
- E.g. Additional factors like early departures and overstays.

- In comparison, other industries like airlines and restaurants have less complications & more flexibility.
- E.g. An airline passenger cannot leave a plane mid-flight.
- E.g. Restaurants can rearrange their interior to create more seating.

Problem Evaluation



References

Chen, L. (2020). Determine the cost of denying boarding to passengers: An optimization-based approach. *International Journal of Advances in Science, Engineering and Technology*, *220*. https://doi.org/International Journal of Production Economics

Hargreaves, C. A., & Tan, S. L. (2020). A Machine Learning Overbooking Algorithm for Enhancing Clinic Efficiency. *International Journal of Advances in Science, Engineering and Technology*, *8*(1), 14-19. https://iraj.doionline.org/dx/IJASEAT-IRAJ-DOIONLINE-16855

Toh, R. (1975). Airline Overbooking Model for Terminator Flights. *International Journal of Transport Economics*, *2*(3), 241-251. https://www.jstor.org/stable/45401705

Toh, R., & Dekay, F. (2002). Hotel Room-inventory Management: An Overbooking Model. *Cornell Hotel and Restaurant Administration Quarterly*, *43*(4), 79-90. https://doi-org.libproxy.smu.edu.sg/10.1177/0010880402434008