PROJECT

DATA223 - ISP

HOTEL BOOKING DATA

PRESENTED BY AADYA GUPTA

COLUMNS AND MEANING

hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	children	babies	meal
Resort Hotel	0	342	2015	July	27	1	0	0	2	0	0	BB
Resort Hotel	0	737	2015	July	27	1	0	0	2	0	0	BB
Resort Hotel	0	7	2015	July	27	1	0	1	1	0	0	BB
Resort Hotel	0	13	2015	July	27	1	0	1	1	0	0	BB
Resort Hotel	0	14	2015	July	27	1	0	2	2	0	0	BB

country	market_segment	distribution_channel	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	reserved_room_type	assigned_room_type	booking_changes	deposit_type	agent	company
PRT	Direct	Direct	0	0	0	C	C	3	No Deposit	NULL	NULL
PRT	Direct	Direct	0	0	0	C	C	4	No Deposit	NULL	NULL
GBR	Direct	Direct	0	0	0	A	C	0	No Deposit	NULL	NULL
GBR	Corporate	Corporate	0	0	0	A	A	0	No Deposit	304	NULL
GBR	Online TA	TA/TO	0	0	0	A	A	0	No Deposit	240	NULL

days_in_waiting_list	customer_type	adr	required_car_parking_spaces	total_of_special_requests	reservation_status	reservation_status_date
0	Transient	0.00	0	0	Check-Out	2015-07-01
0	Transient	0.00	0	0	Check-Out	2015-07-01
0	Transient	75.00	0	0	Check-Out	2015-07-02
0	Transient	75.00	0	0	Check-Out	2015-07-02
0	Transient	98.00	0	1	Check-Out	2015-07-03

COLUMNS AND MEANING

- **hotel** – Type of hotel (e.g., Resort Hotel or City Hotel).
- **is_canceled** – 1 if the booking was canceled, 0 if not.
- **lead_time** – Number of days between the booking date and the arrival date.
- **arrival_date_year** – Year of arrival.
- **arrival_date_month** – Month of arrival.
- **arrival_date_week_number** – Week number of the arrival date.
- **arrival_date_day_of_month** – Day of the month for arrival
- **stays_in_weekend_nights** – Number of weekend nights (Saturday or Sunday).
- **stays_in_week_nights** – Number of weekday nights (Monday-Friday).
- **adults** – Number of adults.
- **children** – Number of children.
- **babies** – Number of babies
- **meal** – Type of meal (BB, HB, FB).
- **country** – Country of origin of the guest.
- **market_segment** – Market segment that made the booking (Direct, Corporate, Online TA, etc.).
- **distribution_channel** – Distribution channel (Direct, TA/TO, etc.).
- **is_repeated_guest** – 1 if a repeated guest, 0 otherwise.
- **previous_cancellations** – Number of previous cancellations.
- **previous_bookings_not_canceled** – Number of previous bookings that were not canceled.
- **reserved_room_type** – Code of the reserved room type.
- **assigned_room_type** – Code of the assigned room type.
- **booking_changes** – Number of changes made to the booking.
- **deposit_type** – Type of deposit (No Deposit, Non Refund, etc.).
- **agent** – ID of the travel agent that made the booking.
- **company** – ID of the company that made the booking.
- **days_in_waiting_list** – Number of days the booking was on the waiting list.
- **customer_type** – Type of customer (Transient, Group, etc.).
- **adr** – Average Daily Rate (ADR), i.e., revenue per available room.
- **required_car_parking_spaces** – Number of required parking spaces.
- **total_of_special_requests** – Number of special requests made by the guest.
- **reservation_status** – Status of the reservation (Check-Out, Canceled, etc.).
- **reservation_status_date** – Date when the last status was updated.

DATA SUMMARY

- **Source & Structure**
  - Origin: Portugal
  - 119,390 rows × 32 columns
  - Mixed variable types (numeric, categorical, dates)

- **Variable: is_canceled**
  - 0 = Not Canceled, 1 = Canceled
  - 37% of bookings were canceled

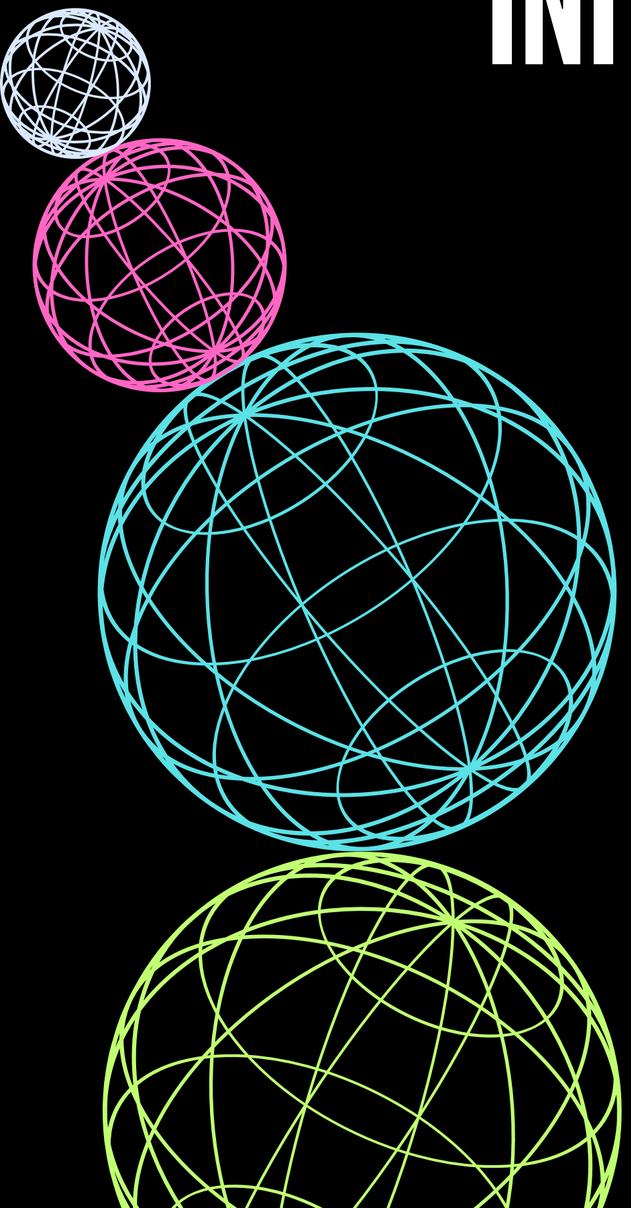
- **Time-related Features**
  - arrival_date_year: 2015 to 2017
  - lead_time: Up to 737 days in advance

- **Cancellations & Bookings:**
  - ~37% bookings were canceled (is_canceled mean = 0.3704).
  - Some bookings were made up to 737 days in advance (lead_time max).
  - Most stays had 2 adults; rare outliers include up to 55 adults.

- **Categorical Variables**
  - hotel, meal, market_segment, distribution_channel, country, etc.
  - Agent and Company: Contain "NULL" values.

- **Data Quality & Outliers:**
  - Some extreme values (e.g., 10 babies, 5400 ADR) suggest outliers.
  - children has 4 missing values; agent & company have "NULL" strings.
  - adr (average daily rate) has a few unusual negatives (min = -6.38).

INITIAL DATA CLEANING



- Dropped unnecessary columns (reservation_status_date, agent, company):

→ Too granular or many NAs — not needed for high-level analysis.

- Created a single arrival_date column:

→ Easier for trend analysis & plotting.

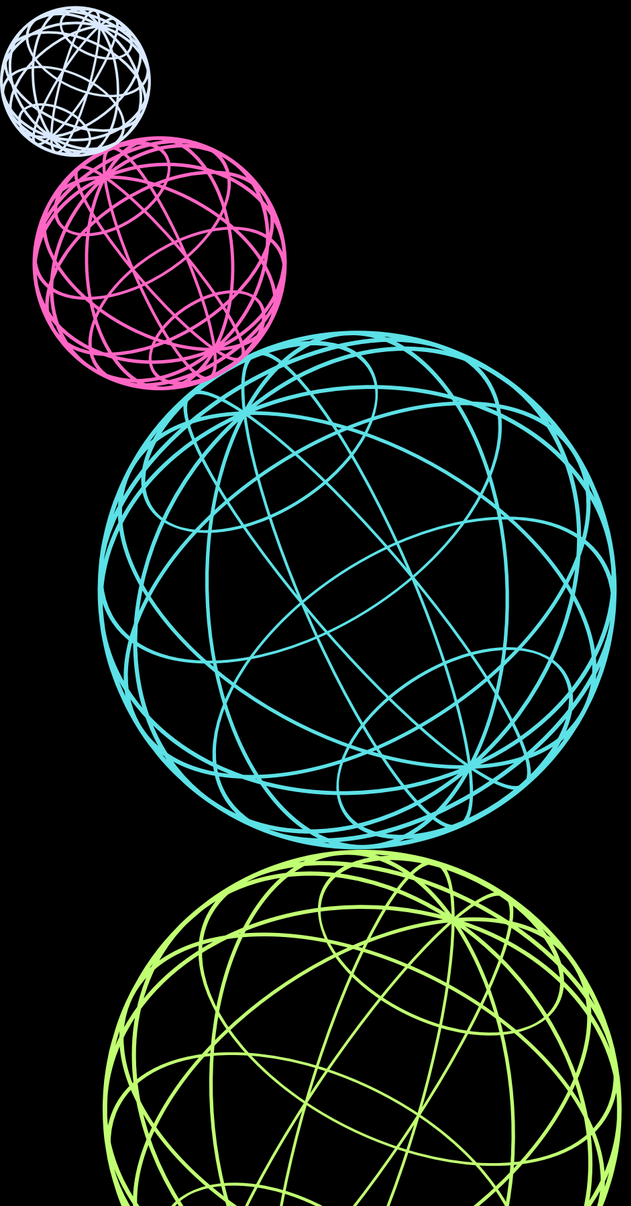
- Dropped original date parts (year, month, day, week_number):

→ Redundant after creating arrival_date.

- Grouped rare countries (<50 bookings) and NA as "Other":

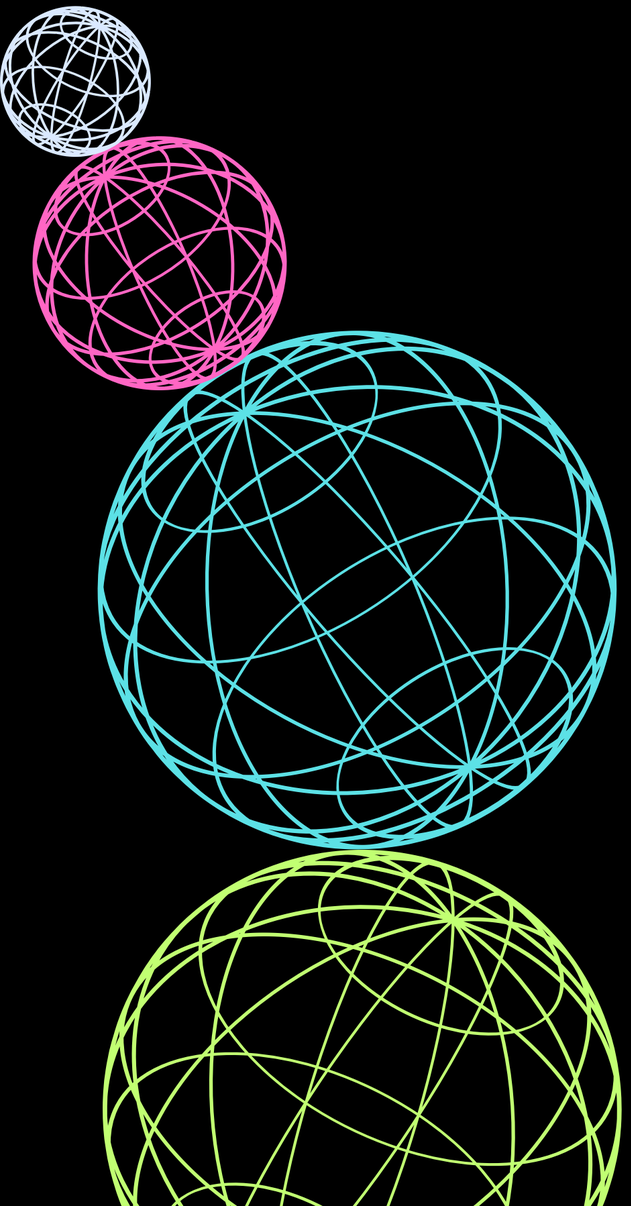
→ Makes visuals cleaner and analysis more focused.

INITIAL DATA CLEANING



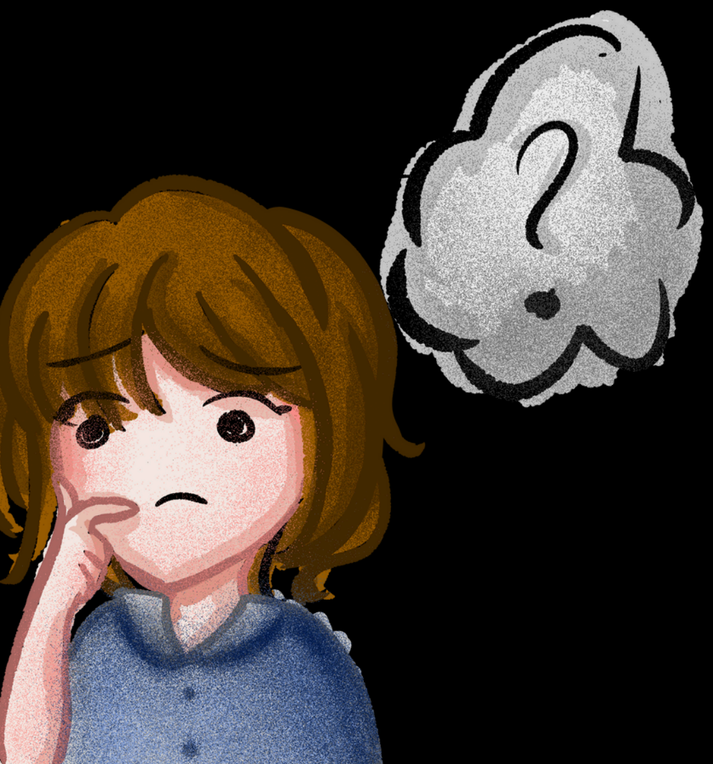
Column Added	Why?
stay_duration = weekend + week nights	Needed for Q1 and Q6 (stay duration vs. family size or special requests)
total_guests = adults + children + babies	Useful for family size analysis in Q6
revenue = adr * stay_duration only if not cancelled	Need this in Q2 for revenue impact of cancellations
Recode meal values into full forms	Improves clarity for Q4 (effect of meal packages)
Recode customer_type into simpler, readable forms	Makes grouping easier, especially if exploring behavior of walk-ins vs. corporates

INITIAL DATA CLEANING



	stay_duration	total_guests	revenue	meal	customer_type
1	0	2	0.00	Breakfast Only	Walk-in
2	0	2	0.00	Breakfast Only	Walk-in
3	1	1	75.00	Breakfast Only	Walk-in
4	1	1	75.00	Breakfast Only	Walk-in
5	2	2	196.00	Breakfast Only	Walk-in
6	2	2	196.00	Breakfast Only	Walk-in
7	2	2	214.00	Breakfast Only	Walk-in
8	2	2	206.00	Full Board	Walk-in
9	3	2	0.00	Breakfast Only	Walk-in
10	3	2	0.00	Half Board	Walk-in
11	4	2	0.00	Breakfast Only	Walk-in
12	4	2	580.00	Half Board	Walk-in
13	4	2	388.00	Breakfast Only	Walk-in
14	4	3	619.08	Half Board	Walk-in
15	4	2	378.84	Breakfast Only	Walk-in

QUESTIONS ?



1. How do special requests and assigned room types influence guest satisfaction and stay duration?

2. How do lead time and booking cancellations impact revenue?

3. How does the country of origin influence booking behavior and lead time?

4. Does the presence of meal packages affect booking modifications or cancellations?

5. How do waiting list bookings and repeated guest status affect stay modifications and cancellations?

1. HOW DO SPECIAL REQUESTS AND ASSIGNED ROOM TYPES INFLUENCE GUEST SATISFACTION AND STAY DURATION?

What I expected (Hypothesis):

- More special requests → might indicate more engaged guests → possibly longer stay or higher satisfaction.
- Assigned room type ≠ reserved room type → might lead to dissatisfaction or shorter stays.
- Guests who get what they booked and have requests fulfilled likely have a better experience.

Variables involved:

- total_of_special_requests
- reserved_room_type
- assigned_room_type
- stay_duration (already created)
- is_canceled

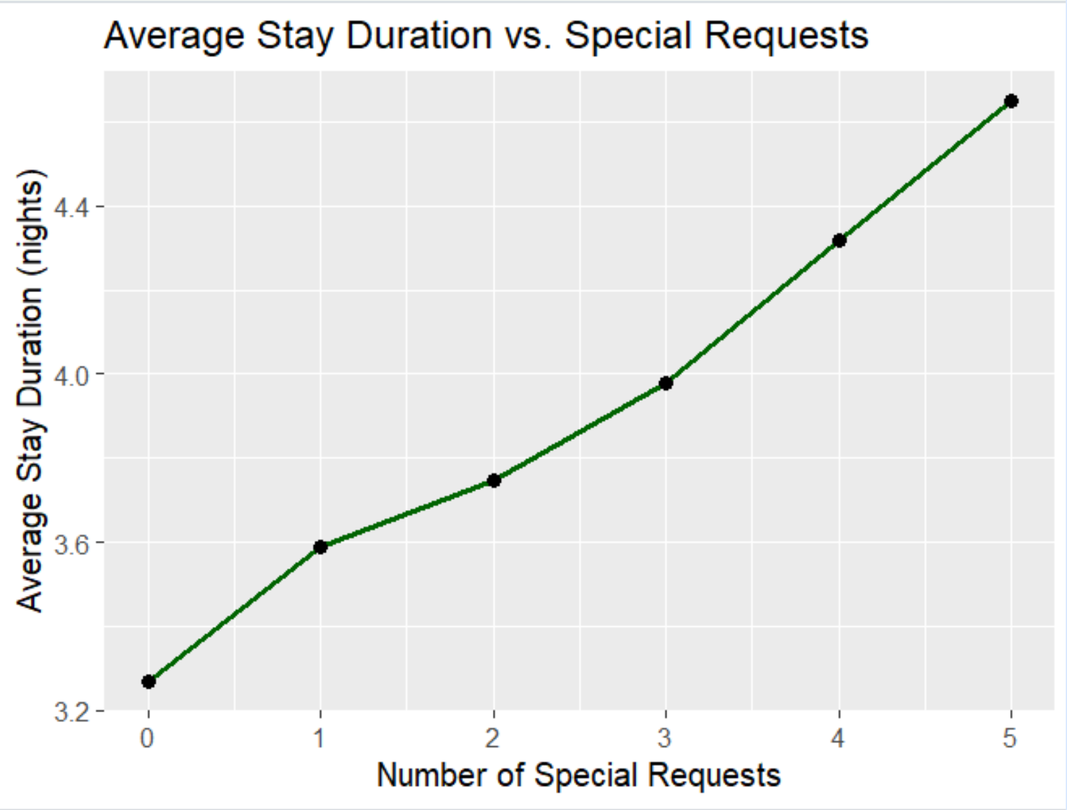
What I Checked	Why It Matters
Stay Duration vs. Special Requests	Are engaged guests staying longer?
Stay Duration by Room Match	Does getting the booked room build trust?
Cancellation Rate vs. Special Requests	Are special request guests more committed?
Cancellation Rate by Room Match	Does mismatch cause dissatisfaction?

STEP

1. Line Plot – Stay Duration and Special Requests

- **Assumption**: More special requests indicate a more invested or excited guest, likely to stay longer.
- **Insight**: Avg stay rises from 3.27 nights (0 requests) to 4.65 nights (5 requests).
- **Analysis**: Guests with more requests may be planning special occasions or have specific needs, leading to longer, more meaningful stays. This hints at **higher satisfaction and commitment.**
- Therefore, **assumption matches the insights found.**

```
> # Avg stay duration for each number of special requests
> stay_by_requests <- hotel_data %>%
+   group_by(total_of_special_requests) %>%
+   summarise(avg_stay = mean(stay_duration))
> # View numeric output
> print(stay_by_requests)
# A tibble: 6 × 2
  total_of_special_requests avg_stay
                      <int>    <dbl>
1                         0     3.27
2                         1     3.59
3                         2     3.75
4                         3     3.98
5                         4     4.32
6                         5     4.65
```

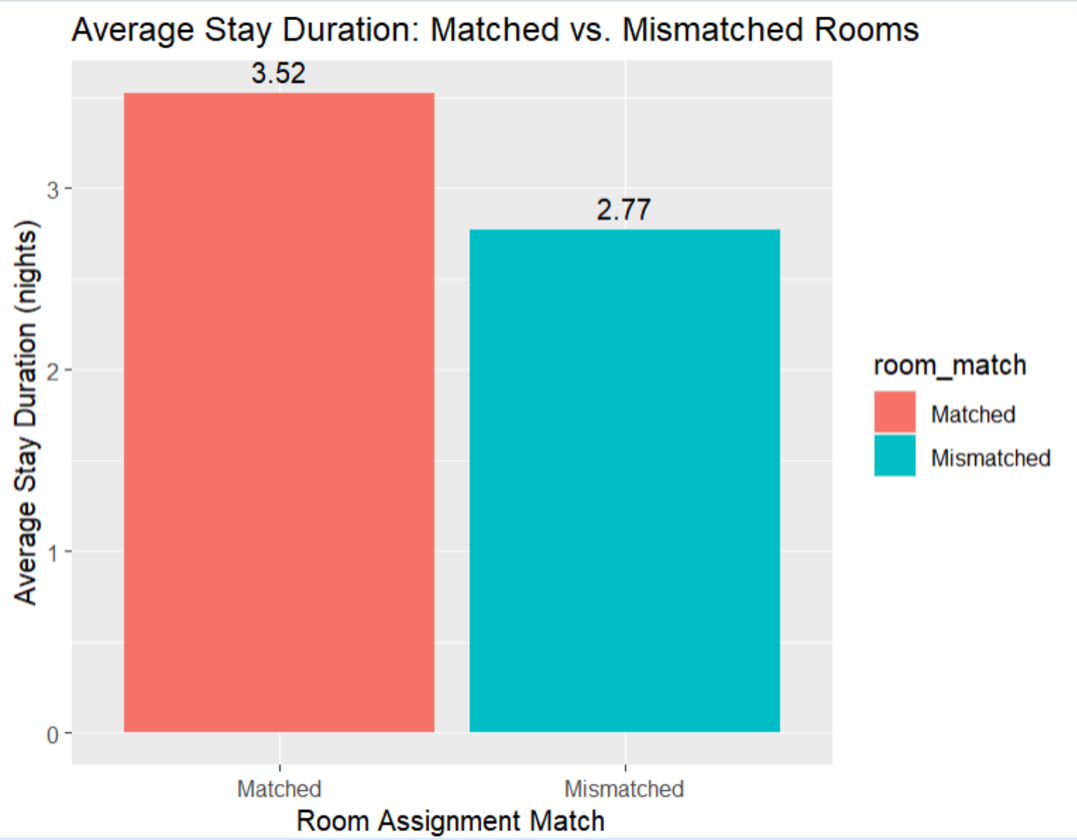


STEP

2. Bar Chart – Stay Duration by Room Match

- **Assumption**: Guests are happier when they get the room they expected.
- **Insight**: Matched room guests stay slightly longer on average (3.52 days) compared to mismatched (2.77 days)
- **Analysis**: Getting what you book builds trust. A matched room assignment likely results in a smoother check-in and fewer surprises, reflecting positive guest experience and a possible willingness to extend or repeat stays.
- Therefore, **assumption matches the insights found.**

```
> stay_by_room <- hotel_data %>%
+   group_by(room_match) %>%
+   summarise(avg_stay = mean(stay_duration))
> # View numeric output
> print(stay_by_room)
# A tibble: 2 × 2
  room_match avg_stay
  <chr>          <dbl>
1 Matched         3.52
2 Mismatched      2.77
```

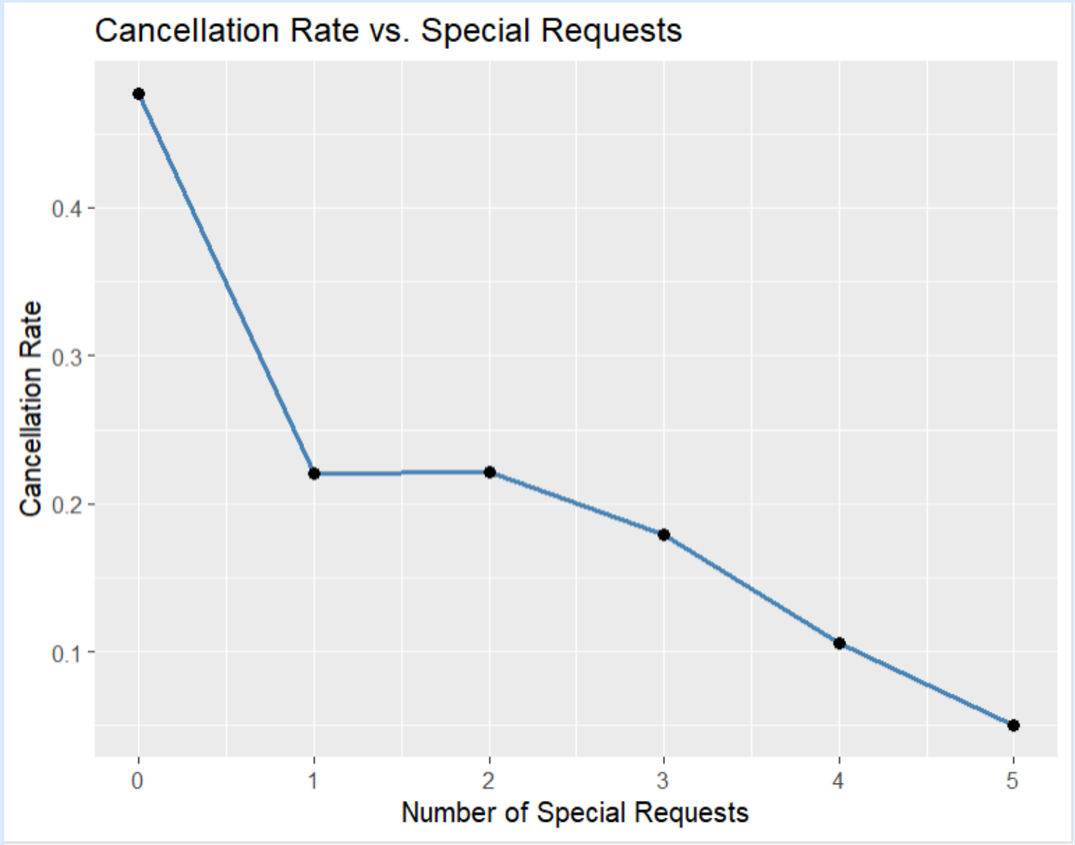


STEP

3. Bar Chart – Cancellation Rate by Special Requests

- **Assumption**: Guests making more requests are more likely to follow through with their booking.
- **Insight**: Cancellation drops dramatically—from 47.7% (0 requests) to 5% (5 requests).
- **Analysis**: Special requests signal intent and planning. These guests are less impulsive and more committed, suggesting that accommodating special requests can lead to higher retention and lower cancellation rates.
- Therefore, **assumption matches the insights found.**

```
> cancellation_by_requests <- hotel_data %>%
+   group_by(total_of_special_requests) %>%
+   summarise(cancel_rate = mean(is_canceled))
> # View numeric output
> print(cancellation_by_requests)
# A tibble: 6 × 2
  total_of_special_requests cancel_rate
                      <int>       <dbl>
1                         0       0.477
2                         1       0.220
3                         2       0.221
4                         3       0.179
5                         4       0.106
6                         5       0.05
```

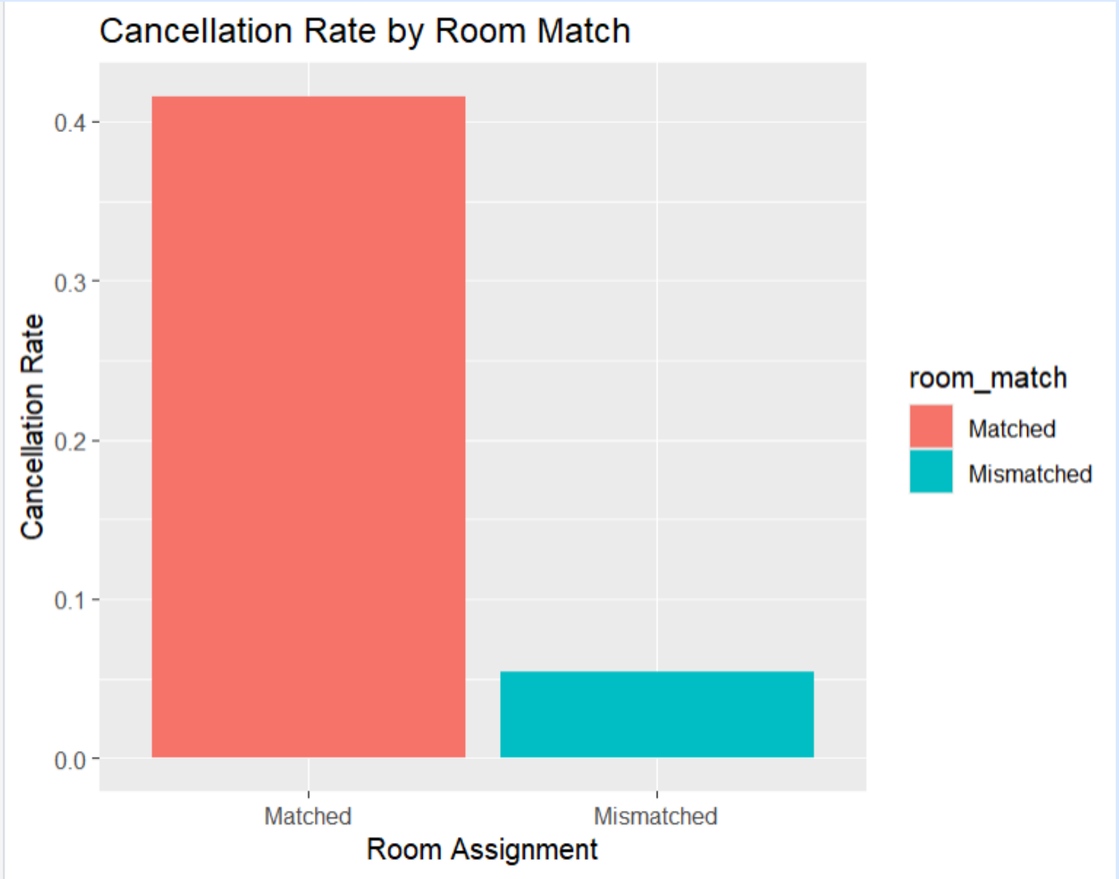


STEP

4. Bar Chart – Cancellation Rate by Room Match

- **Assumption**: Room mismatches lead to dissatisfaction and possibly cancellations.
- **Insight**: Opposite of expected – mismatched rooms have a lower cancellation rate (~5.3%) than matched (~41.6%)
- **Analysis**: This is a data bias. Cancellations typically occur before room assignment. Mismatched rooms are only recorded after check-in, meaning only non-canceled guests can even be marked as mismatched. Hence, this chart doesn't reflect dissatisfaction – it reflects the process of booking and fulfillment.
- Therefore, **assumption does not match the insights found.**

```
> hotel_data %>%
+   group_by(room_match) %>%
+   summarise(cancellation_rate = mean(is_canceled))
# A tibble: 2 × 2
  room_match cancellation_rate
  <chr>                  <dbl>
1 Matched                0.416
2 Mismatched             0.0538
```



2. HOW DO LEAD TIME AND BOOKING CANCELLATIONS IMPACT REVENUE?

What we expect (Hypothesis):

- Longer lead times may associate with higher cancellation rates — people booking far in advance might be less committed or face changing plans.
- Canceled bookings contribute no revenue, so high cancellation rates = potential revenue loss.
- Among non-canceled bookings, longer lead time might mean more planned stays, possibly with higher revenue (e.g., vacation trips or group bookings).

Variables involved:

- lead_time
- is_canceled
- adr (Average Daily Rate)
- stay_duration (already created)
- revenue = adr * stay_duration

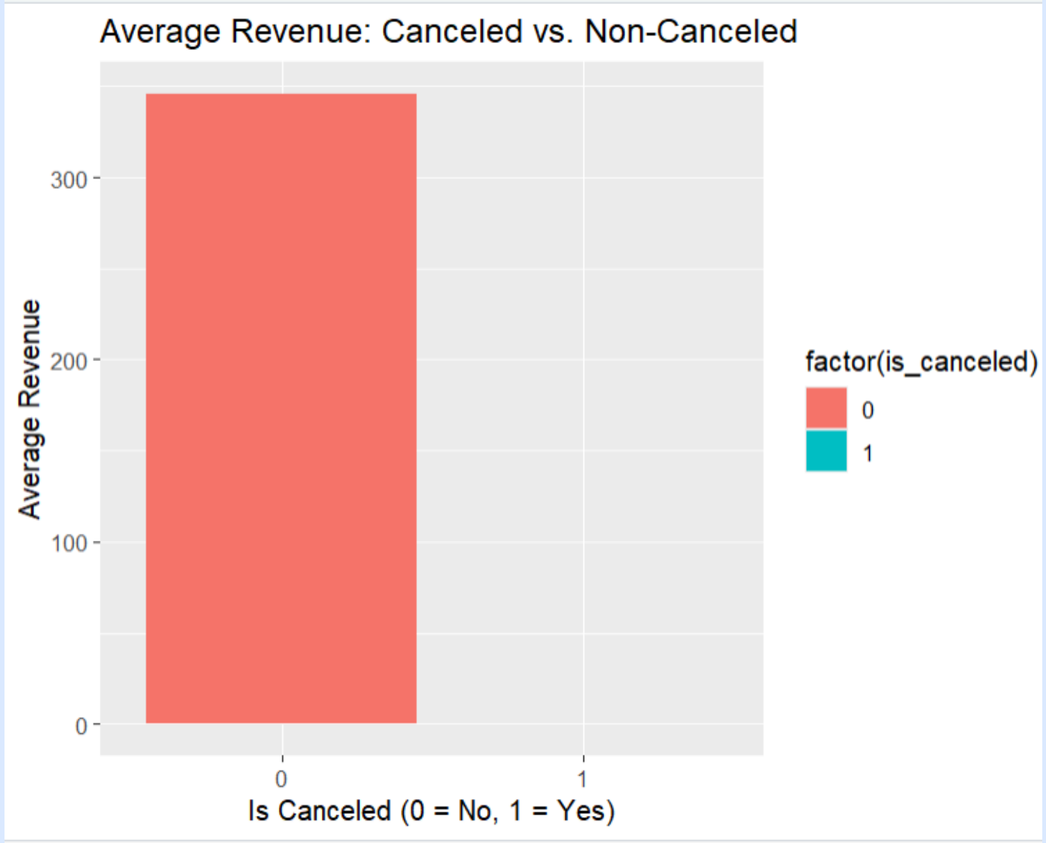
What I Checked	Why It Matters
Compared revenue for canceled vs. non-canceled bookings	Helps quantify revenue loss due to cancellations.
Plotted cancellation rate across lead time bins	Helps understand booking commitment based on planning time.
Analyzed average revenue by lead time (non-canceled only)	Helps identify which booking windows are most profitable.

STEP

1. Bar Chart – Average Revenue: Canceled and Non-Canceled Bookings

- **Assumption**: Canceled bookings = ₹0 revenue. Non-canceled bookings should generate meaningful revenue.
- **Insight**: Canceled bookings averaged ₹0, while non-canceled bookings averaged ₹346.
- **Analysis**: This confirms that if we reduce cancellations, we could boost revenue per booking. We often assume that canceled bookings bring in no money — but I thought of checking with data, because often weird exceptions do show up (e.g., non-refundable fees).
- Therefore, **assumption matches the insights found.**

```
> avg_revenue
# A tibble: 2 × 2
  is_canceled avg_rev
        <int>   <dbl>
1           0    346.
2           1      0
```

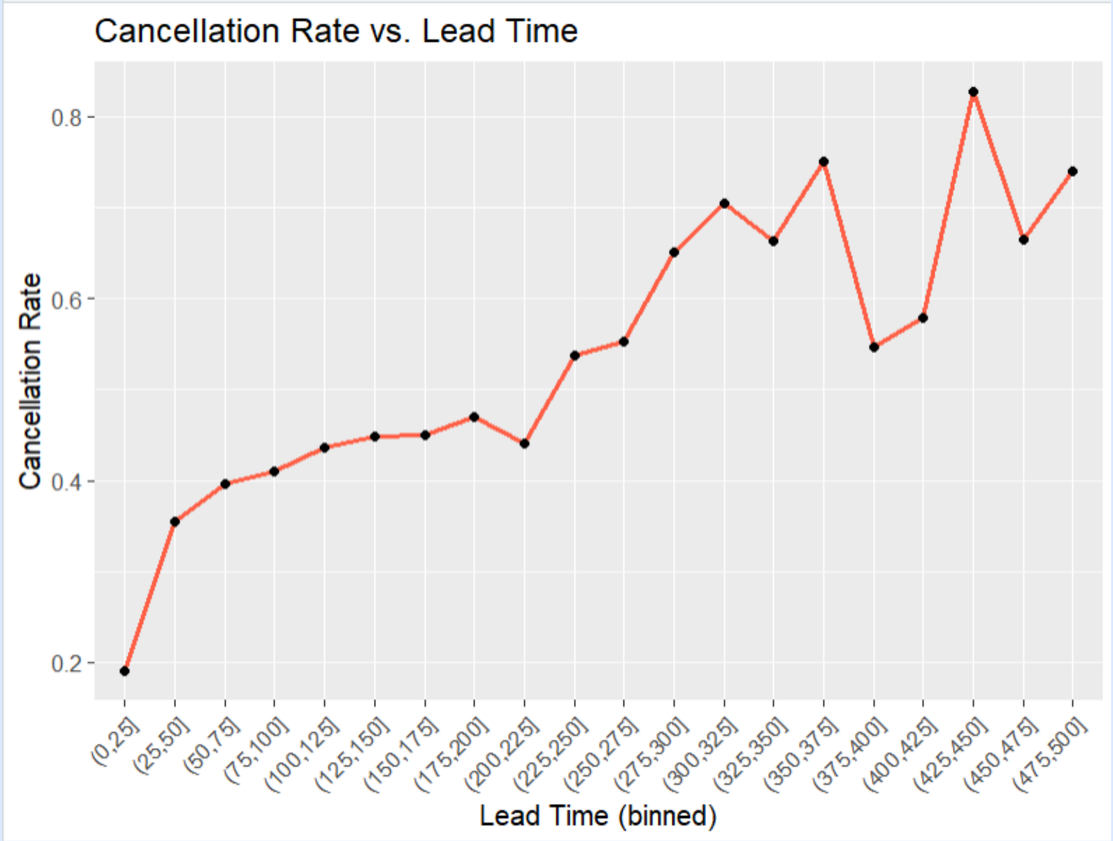


STEP

2. Line Chart – Cancellation Rate by Lead Time (Binned)

- **Assumption**: The longer the lead time, the higher the chance of cancellation.
- **Insight**: Cancellation rate rises steadily with lead time — from ~21% for 0–25 days to over 80% for bookings with lead time > 400 days.
- **Analysis**: This suggests guests with very early bookings are more uncertain or flexible — like booking in case of travel plans or offers. Identifying this trend is important because it allows targeted policies: e.g., stricter cancellation terms or deposits for long lead times to protect revenue.
- Therefore, **assumption matches the insights found.**

```
> cancel_by_lead <- hotel_data %>%
+   mutate(lead_time_bin = cut(lead_time, breaks = seq(0, 500, 25))) %>%
+   filter(!is.na(lead_time_bin)) %>%  # remove NAs created by cut()
+   group_by(lead_time_bin) %>%
+   summarise(cancellation_rate = mean(is_canceled, na.rm = TRUE))
> cancel_by_lead
# A tibble: 20 × 2
   lead_time_bin cancellation_rate
   <fct>                     <dbl>
 1 (0,25]                    0.190
 2 (25,50]                   0.355
 3 (50,75]                   0.397
 4 (75,100]                  0.410
 5 (100,125]                 0.436
 6 (125,150]                 0.448
```



STEP

3. Line Chart – Average Revenue and Lead Time (Only Non-Canceled Bookings)

- **Assumption**: Mid-range bookings may be better planned and less profitable, while last-minute or very early bookings may bring in more. Longer lead time → lesser revenue.

- **Insight**:

-- Revenue peaks at **lead time of 150–200 days (~₹490–₹500)**, then declines or fluctuates for longer lead times.
-- Revenue drops sharply from ₹455 to ₹245 between 350–375 and 375–400 lead days.
-- **Why** the drop in revenue? **No clue – therefore limitation in this analysis**
-- **Steps tried:**

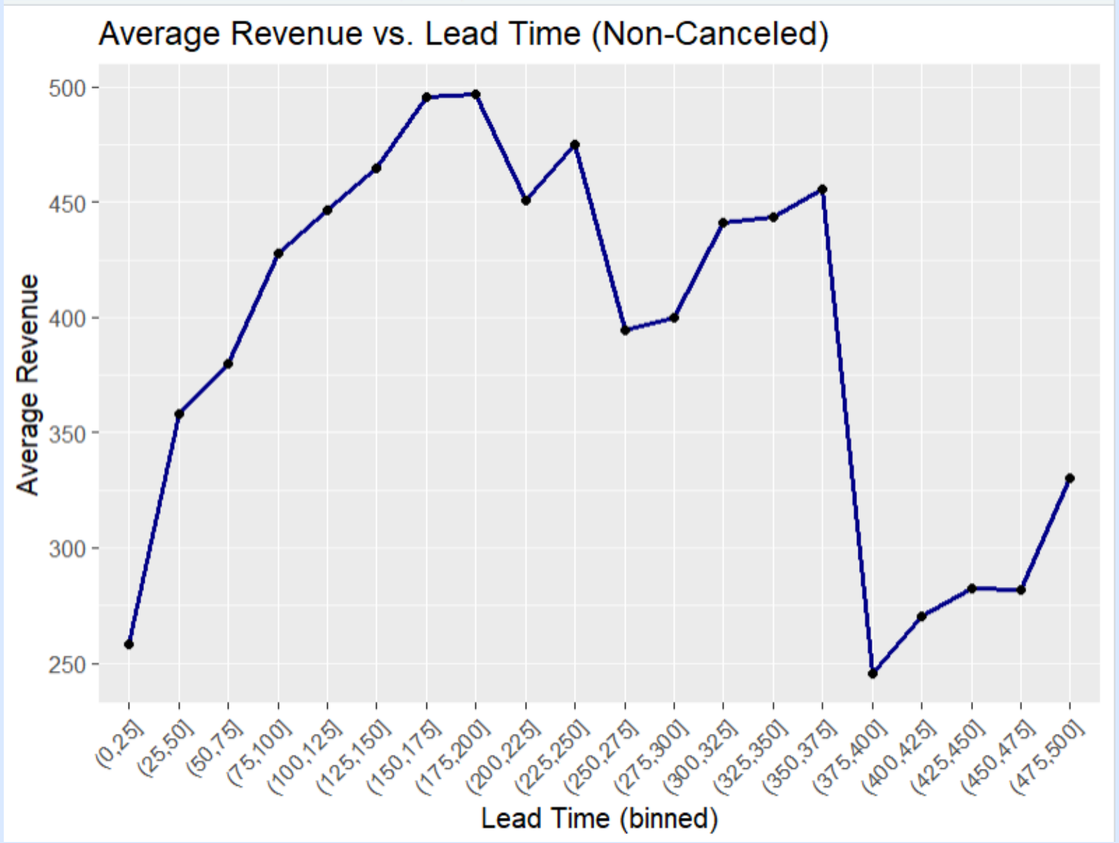
- > Checked sample size → To see if fewer bookings caused the drop, but n = 393, so not a small sample → ❌ Not the reason.
-> Checked avg stay & ADR → To test if lower stay/ADR caused drop → ❌ Both values weren't drastically low → Not the reason.
-> Checked market segment + channel → To see if low-value segments dominated → ❌ Majority were still from TA/TO, like other bins.
-> Compared TA/TO bookings across bins → To check if TA/TO bookings were unusually high in this bin → ❌ Not really, other bins had even higher.
-> Checked if the drop in revenue was part of a bigger trend across lead time ranges.

It showed that revenue really did drop in the (375,400] bin — but it didn't explain why the drop happened, so I was left with a pattern but no clear reason — making it a limitation.

- **Analysis**

A noticeable drop in revenue was found between lead times of 350–400 days. Multiple checks were conducted — including sample size, average stay duration, ADR, and booking segments — but none explained the dip. The pattern exists, but **no conclusive reason could be identified, making it a LIMITATION of the data**.

```
> lead_rev
# A tibble: 20 × 2
   lead_time_bin avg_revenue
   <fct>               <dbl>
 1 (0,25]              258.
 2 (25,50]             358.
 3 (50,75]             380.
 4 (75,100]            428.
 5 (100,125]           447.
 6 (125,150]           465.
 7 (150,175]           495.
 8 (175,200]           497.
 9 (200,225]           451.
10 (225,250]           475.
11 (250,275]           395.
12 (275,300]           400.
13 (300,325]           441.
14 (325,350]           444.
15 (350,375]           455.
16 (375,400]           245.
17 (400,425]           270.
18 (425,450]           282.
19 (450,475]           282.
20 (475,500]           330.
```



3. HOW DOES THE COUNTRY OF ORIGIN INFLUENCE BOOKING BEHAVIOR AND LEAD TIME?

Hypothesis

- International guests likely have longer lead times due to travel, visa, and planning needs.
- Cancellation rates may be higher in some countries (e.g., due to speculative or refundable bookings).
- Certain countries might bring higher ADRs, longer stays, or more special requests depending on the guest profile (e.g., tourists vs. business travelers).

What I Checked	Why It Matters
Calculated average lead time by country	To test if international guests plan further in advance
Calculated cancellation rate by country	To check if international bookings are more likely to get canceled
Calculated average number of special requests	To see if guests from different countries have varying service expectations
Calculated average number of booking changes	To examine if guests with complex travel plans (often international) change bookings more

COLUMNS AND MEANING

- **The following were the top 10 countries:**

  ISO - > Code Country Name

  PRT -> Portugal
  GBR -> United Kingdom
  FRA -> France
  ESP -> Spain
  DEU -> Germany
  ITA -> Italy
  IRL -> Ireland
  BEL -> Belgium
  BRA -> Brazil
  NLD -> Netherlands

Variables involved:

- country
- lead_time
- is_canceled
- adr
- stay_duration
- revenue

STEP

1. **Average Lead Time by Country**

- **Assumption**: International guests tend to plan their trips further in advance, leading to longer lead times.

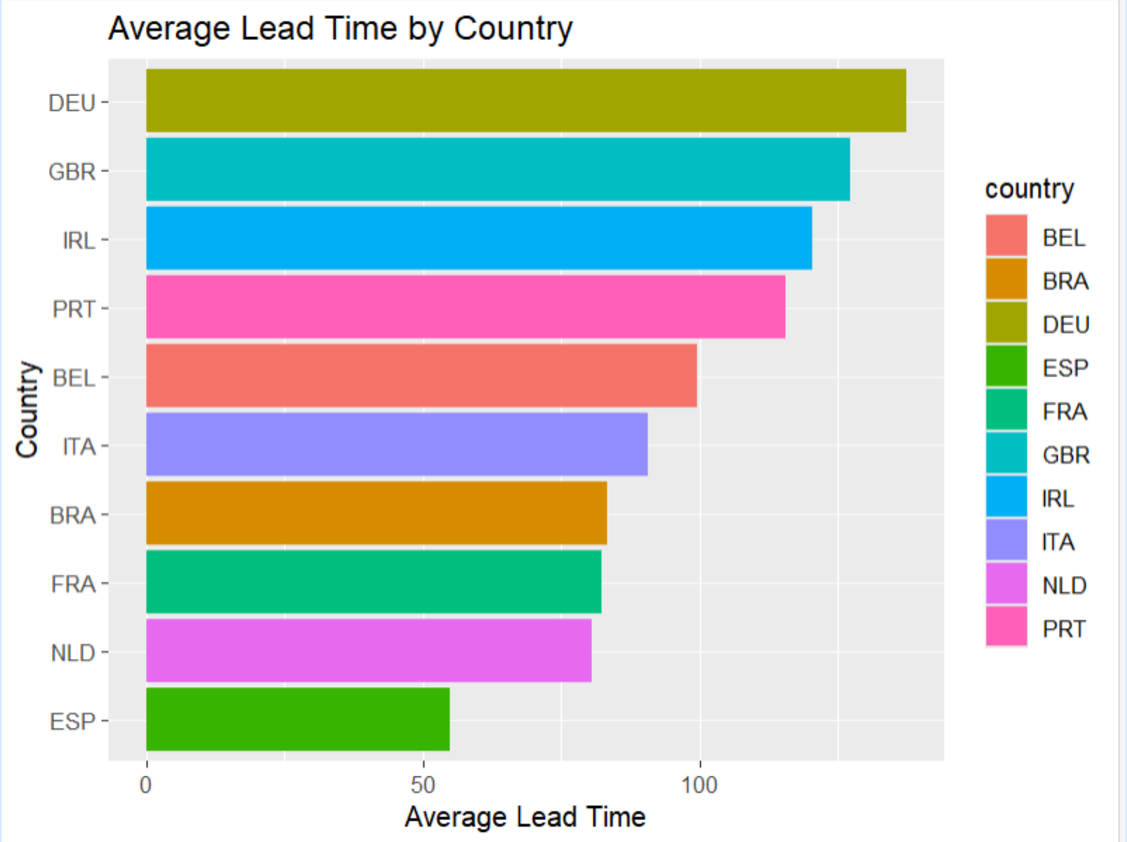
- **Insight**:

- Germany (~137 days), UK (~127 days), and Ireland (~120 days) show the longest lead times.
- Spain (~55 days) and Netherlands (~80 days) book closer to the stay date.

- **Analysis**: Guests from Germany, the UK, and Ireland likely plan their travel well in advance — possibly due to longer travel distances or structured itineraries. Meanwhile, bookings from Spain and the Netherlands seem more spontaneous or last-minute.

- Therefore, **assumption matches the insights found.**

```
> lead_by_country <- hotel_data %>%
+   filter(country %in% top_countries$country) %>%
+   group_by(country) %>%
+   summarise(avg_lead = mean(lead_time, na.rm = TRUE))
> lead_by_country
# A tibble: 10 × 2
   country avg_lead
   <chr>       <dbl>
 1 BEL          99.7
 2 BRA          83.3
 3 DEU         137.
 4 ESP          54.9
 5 FRA          82.3
 6 GBR         127.
 7 IRL         120.
 8 ITA          90.7
 9 NLD          80.6
10 PRT         116.
```



STEP

2. Cancellation Rate by Country

- **Assumption**: International bookings are more uncertain and may have higher cancellation rates

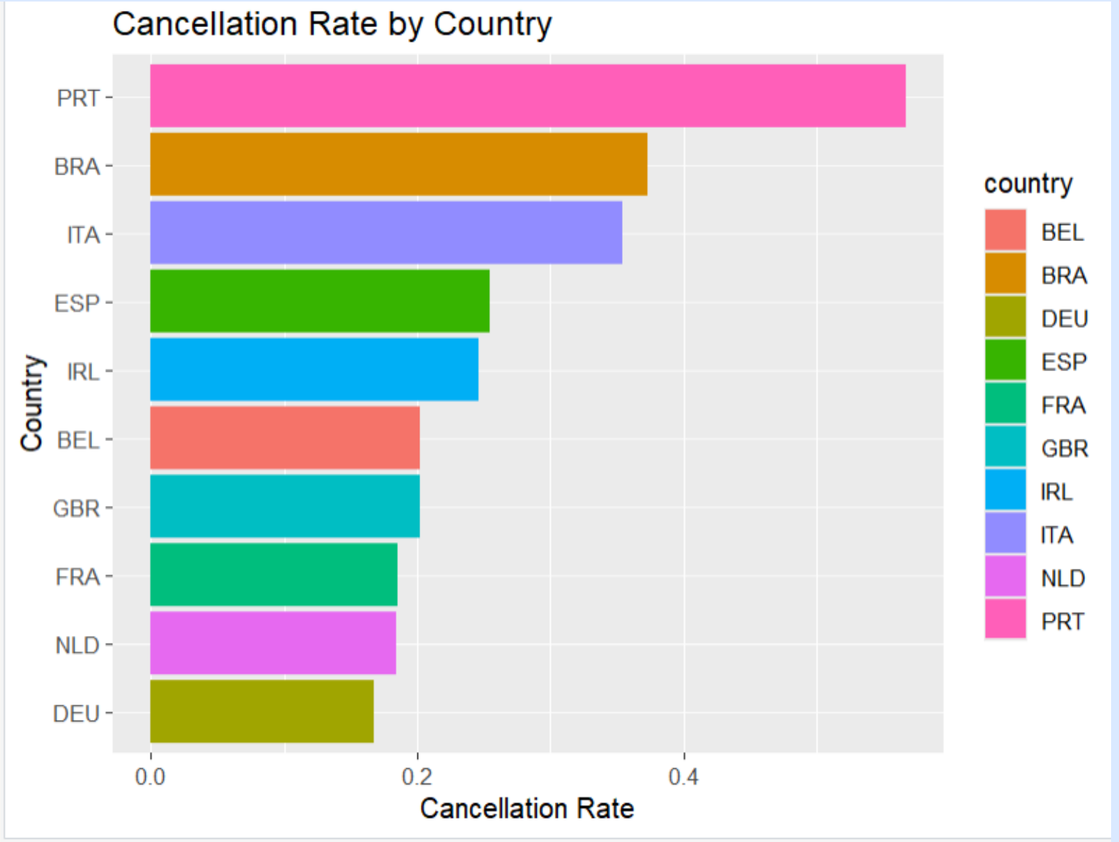
- **Insight**:

- Portugal shows a very high cancellation rate (~56.6%).
- Germany (16.7%), Netherlands (18.4%), and France (18.6%) have much lower rates.

- **Analysis**: The high cancellation rate in Portugal may reflect local behavior — guests booking casually, potentially taking advantage of flexible policies, or making last-minute speculative plans.
- In contrast, German, Dutch, and French guests appear more committed, likely because their travel involves more effort, cost, and planning — reducing the chances of cancellations.
- This suggests that proximity and ease of travel may actually increase cancellation rates due to lower perceived cost or risk in canceling.

- **Therefore, this goes against our assumptions.**

```
> cancel_by_country <- hotel_data %>%
+   filter(country %in% top_countries$country) %>%
+   group_by(country) %>%
+   summarise(cancellation_rate = mean(is_canceled, na.rm = TRUE))
> cancel_by_country
# A tibble: 10 × 2
   country cancellation_rate
   <chr>               <dbl>
 1 BEL                 0.202
 2 BRA                 0.373
 3 DEU                 0.167
 4 ESP                 0.254
 5 FRA                 0.186
 6 GBR                 0.202
 7 IRL                 0.247
 8 ITA                 0.354
 9 NLD                 0.184
10 PRT                 0.566
```



STEP

3. Average Special Requests by Country

- **Assumption**: Guests from different countries may have different service expectations or cultural preferences, leading to variation in how often they make special requests.

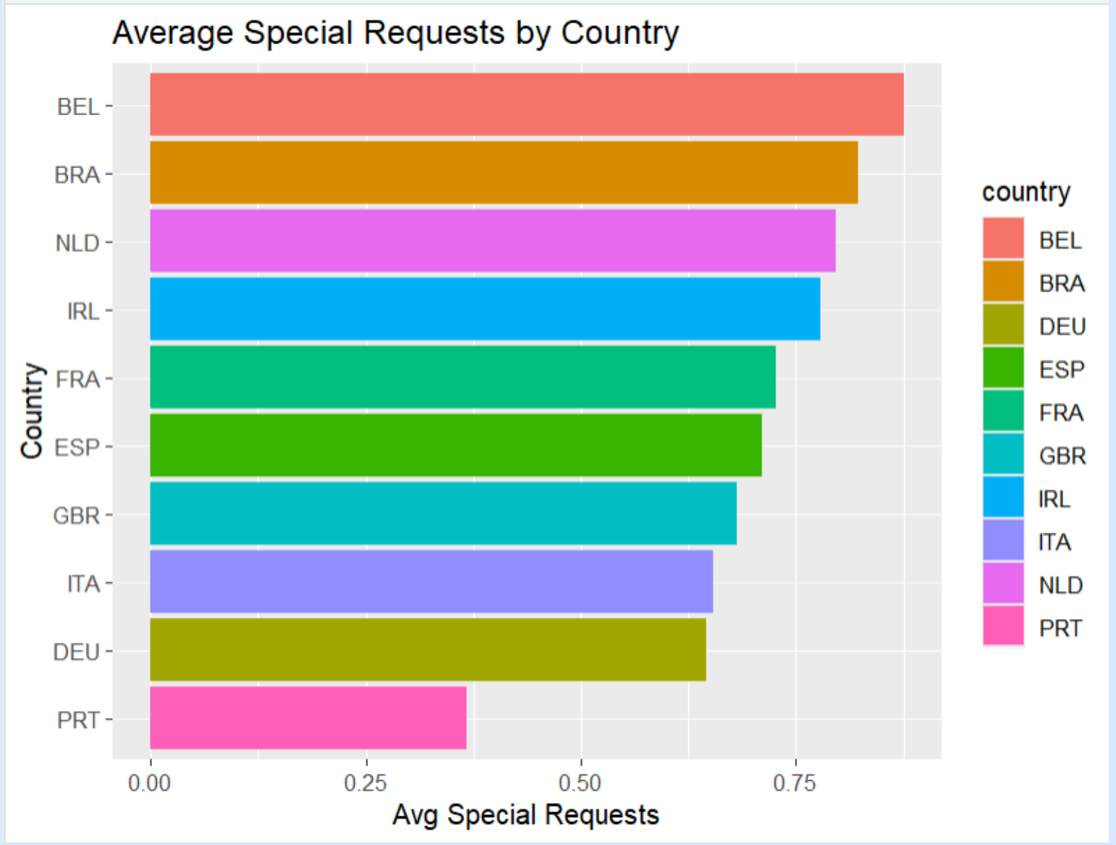
- **Insight**:

- Highest: Belgium (87.5%), Brazil (82.3%).
- Lowest: Portugal (36.7).

- **Analysis**: Guests from Belgium and Brazil may expect more customized service or have more specific needs — potentially reflecting longer stays, family/group travel, or cultural habits around hospitality. Meanwhile, Portuguese guests — being locals — may stay for shorter durations, feel more familiar or comfortable, or simply expect fewer customizations.

- **Therefore, assumption matches the insights found.**

```
> requests_by_country <- hotel_data %>%
+   filter(country %in% top_countries$country) %>%
+   group_by(country) %>%
+   summarise(avg_special_requests = mean(total_of_special_requests, na.rm = TRUE))
> requests_by_country
# A tibble: 10 × 2
   country avg_special_requests
   <chr>                  <dbl>
 1 BEL                    0.875
 2 BRA                    0.823
 3 DEU                    0.645
 4 ESP                    0.710
 5 FRA                    0.727
 6 GBR                    0.681
 7 IRL                    0.779
 8 ITA                    0.654
 9 NLD                    0.797
10 PRT                    0.367
```



STEP

4. Average Booking Changes by Country

- **Assumption**: Guests from farther locations or with more complex travel plans are more likely to make booking changes.

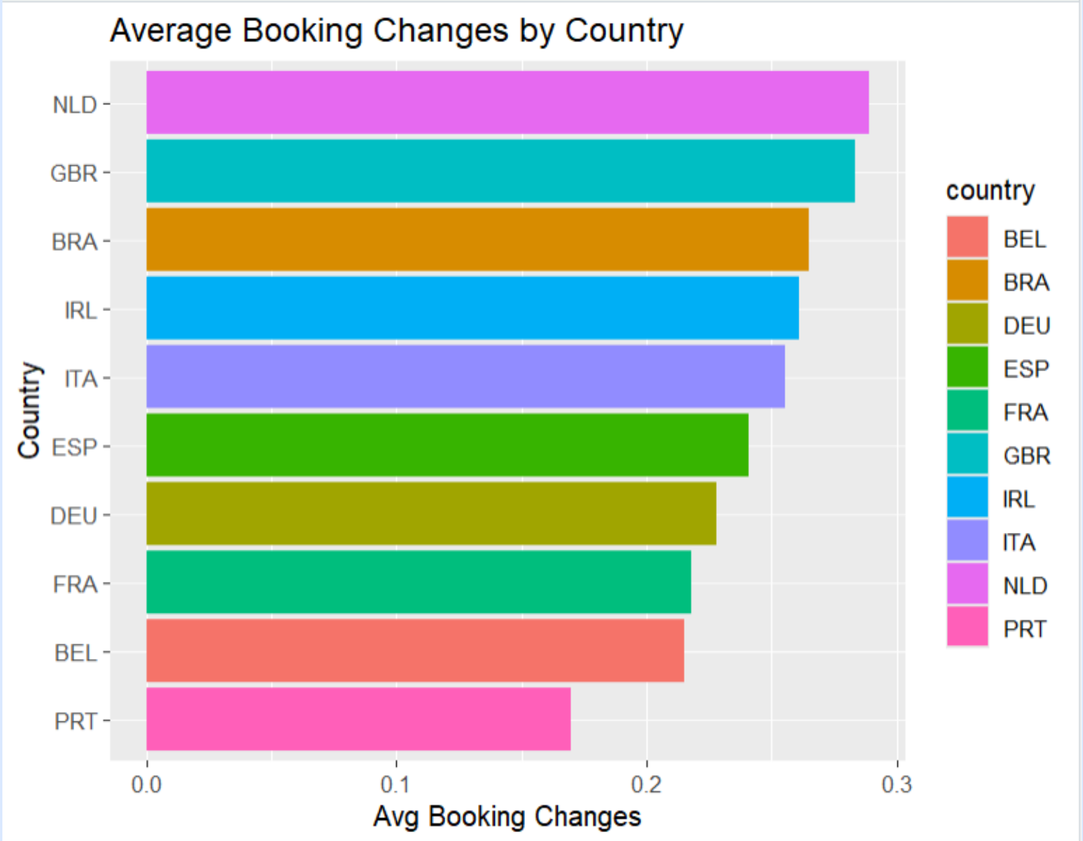
- **Insight**:

- Highest: Netherlands (0.289), UK (0.283).
- Lowest: Portugal (0.170).

- **Analysis**: Guests from the Netherlands and UK tend to modify their bookings more often. This could be because they book early, giving them more time to change details (e.g., dates, number of guests).
- These guests may also be on longer, international trips that require coordination and therefore need updates.
- Portuguese guests, on the other hand, likely book closer to the stay date, or for shorter stays, so there's less opportunity or need to make changes.

- **Therefore, assumption matches the insights found.**

```
> changes_by_country
# A tibble: 10 × 2
   country avg_changes
   <chr>         <dbl>
 1 BEL           0.215
 2 BRA           0.265
 3 DEU           0.228
 4 ESP           0.241
 5 FRA           0.218
 6 GBR           0.283
 7 IRL           0.261
 8 ITA           0.256
 9 NLD           0.289
10 PRT           0.170
```



4. DOES THE PRESENCE OF MEAL PACKAGES AFFECT BOOKING CHANGES OR CANCELLATIONS?

What we expect (Hypothesis):

- Meal packages might reduce cancellations, as guests who book meals are likely more invested in the stay.
- Meal packages may reduce booking changes, since meals suggest well-planned itineraries.
- Alternatively, guests with meal plans may change bookings more if they're adjusting travel plans or customizing their stay.

Variables involved:

- meal
- is_canceled
- booking_changes

What I Checked	Why It Matters
Calculated cancellation rate by meal type	To see if meal packages reduce cancellations
Calculated average booking changes by meal type	To check if guests with meal packages make fewer changes
Calculated booking volume per meal type	To understand which meal types are most popular — giving context to the other metrics

STEP

1. Cancellation Rate by Meal Type

- **Assumption**: Guests with meal packages (especially Full Board or Half Board) are more committed and less likely to cancel.

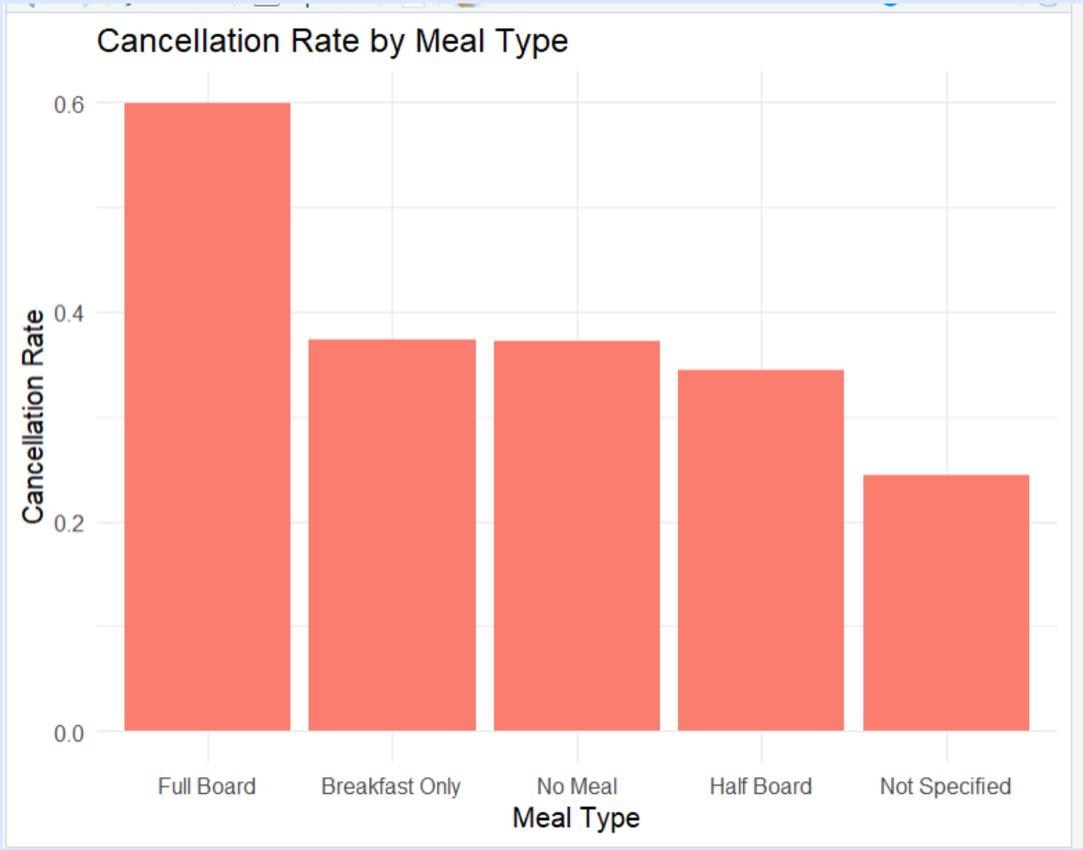
- **Insight**:

- Full Board: 59.9%
- Not Specified: 24.5%

- **Analysis**: Full Board guests canceled the most — possibly because these are high-commitment bookings, and cancellations may reflect uncertain travel plans or bulk tour cancellations. Guests with "Not Specified" meals likely reflect local or business travelers with short, more confirmed stays.

- Therefore, **assumption is different from the insights.**

```
> cancel_rate_meal
# A tibble: 5 × 4
  meal           total_bookings total_cancellations cancellation_rate
  <chr>                   <int>               <int>             <dbl>
1 Breakfast Only          92310               34510             0.374
2 Full Board                798                 478             0.599
3 Half Board              14463                4984             0.345
4 No Meal                 10650                3966             0.372
5 Not Specified            1169                 286             0.245
```



STEP

2. Average Booking Changes by Meal Type

- **Assumption**: Guests with structured packages (like Full Board) will make more changes; those without meals may change less.

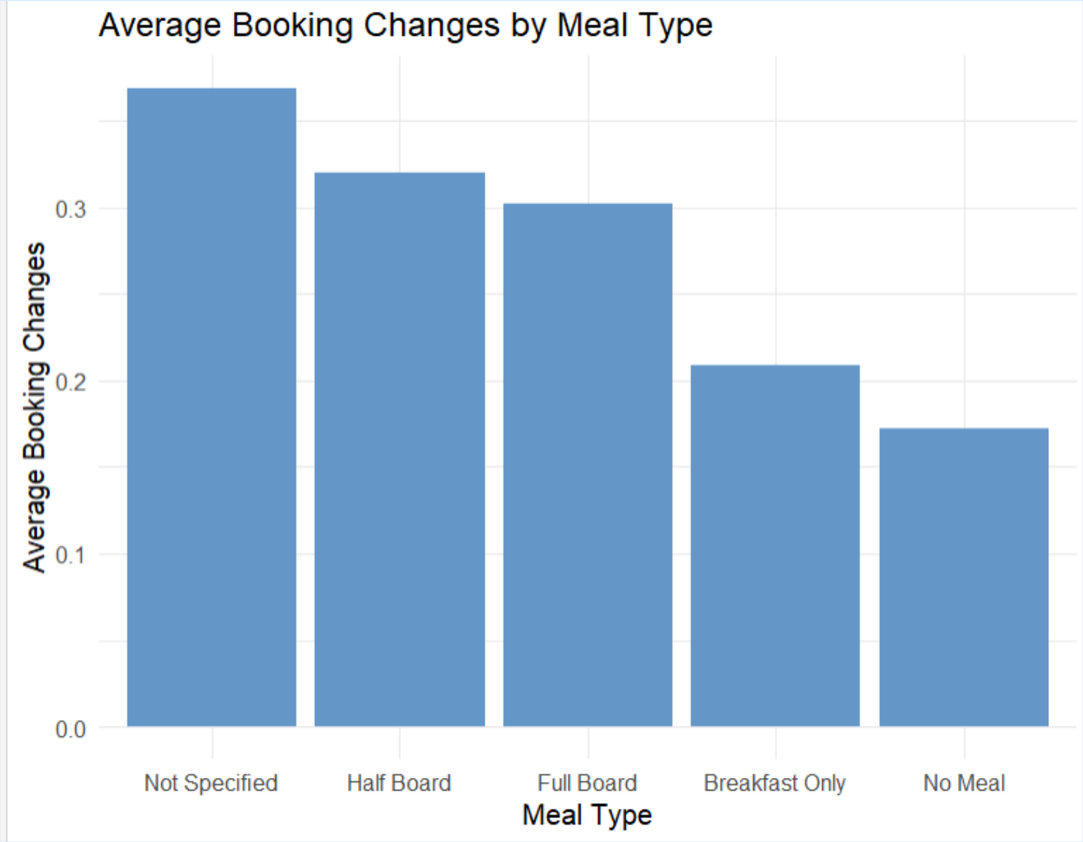
- **Insight**:

- Not Specified: 0.369 (Highest)
- No Meal: 0.172 (Lowest)

- **Analysis**: "Not Specified" and Full/Half Board guests changed their bookings more, possibly because they booked in advance and had more time to modify plans. "No Meal" guests may book last-minute or for simple stays, leading to fewer changes.

- Therefore, **assumption matches insights**

```
> booking_changes_meal <- hotel_data %>%
+   group_by(meal) %>%
+   summarise(
+     avg_changes = mean(booking_changes, na.rm = TRUE))
> booking_changes_meal
# A tibble: 5 × 2
  meal            avg_changes
  <chr>                 <dbl>
1 Breakfast Only        0.209
2 Full Board            0.302
3 Half Board            0.320
4 No Meal               0.172
5 Not Specified         0.369
```



STEP

3. Meal Type Popularity (Booking Volume)

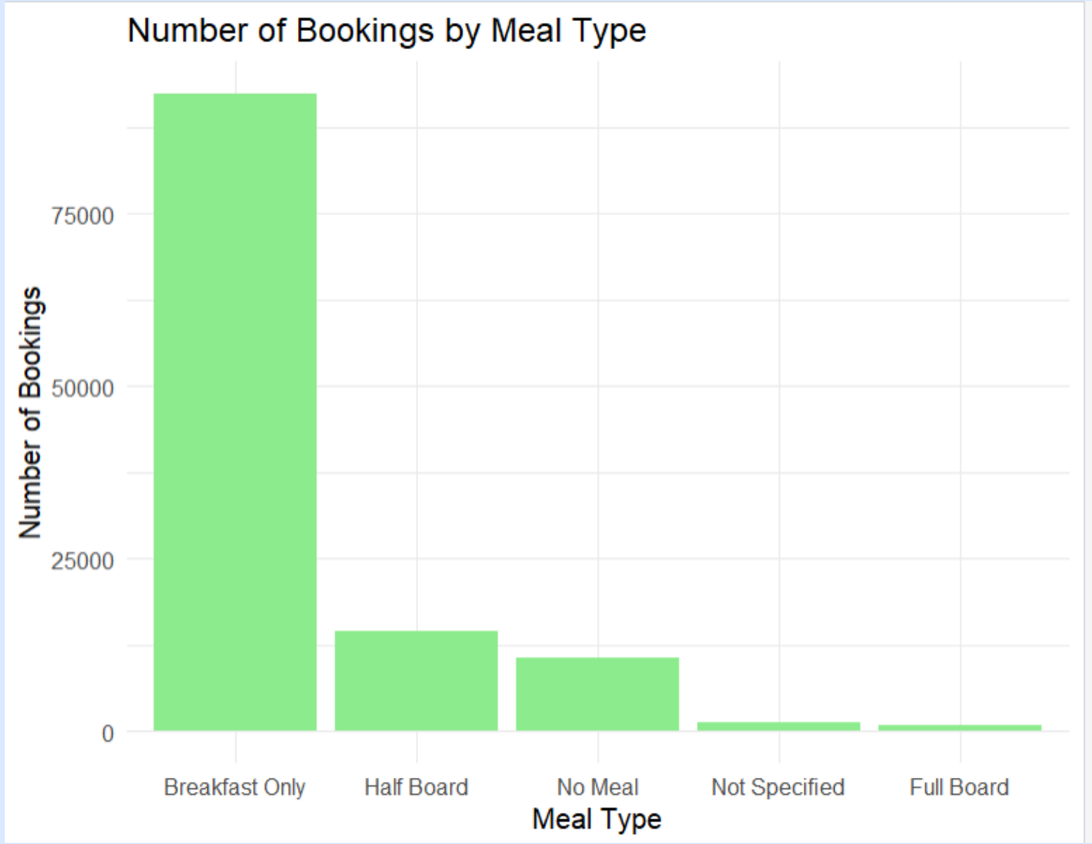
- **Why**: Understanding booking volume helps prioritize which meal types matter most for business decisions.

- **Insight**:

- Breakfast Only: 92,310 bookings 🍽️ (Most popular)
- Full Board: 798 bookings (Least popular)

- **Analysis**: "Breakfast Only" dominates the booking volume — likely due to being the default or most affordable option in many hotel packages.
- "Full Board" is the least common, which may be because it appeals to a niche audience (e.g., tourists on long vacations or package deals).
- Most travelers prefer some flexibility — breakfast included, but other meals outside the hotel — especially in cities with diverse dining options.

```
> meal_counts <- hotel_data %>%
+   count(meal)
> meal_counts
            meal     n
1 Breakfast Only 92310
2      Full Board   798
3      Half Board 14463
4         No Meal 10650
5  Not Specified  1169
```



5. HOW DO WAITING LIST BOOKINGS AND REPEATED GUEST STATUS AFFECT STAY MODIFICATIONS AND CANCELLATIONS?

What we expect (Hypothesis):

- Repeated guests are more likely to stick to their plans, so they may cancel less and change bookings less frequently.
- Guests on the waiting list might have higher cancellation or booking change rates, since their stay is uncertain to begin with.

Variables involved:

- is_repeated_guest
- is_canceled
- booking_changes
- days_in_waiting_list
- waitlist_status

- "Waitlisted": if days_in_waiting_list > 0
- "Not Waitlisted": otherwise

What I Checked	Why It Matters
cancellation rate by repeated guest status	To see if loyalty associates with commitment
average booking changes by repeated guest status	To assess if familiarity leads to fewer modifications
cancellation rate by waitlist status	To check if uncertainty increases cancellations
average booking changes by waitlist status	To understand if waitlisted guests engage less with booking adjustments

STEP

1. Cancellation Rate by Repeated Guest Status

Assumption: Repeated guests are more committed and likely to follow through with their stays.

- **Insight**:

- New Guests (0): 37.8%
- Repeated Guests (1): 14.5%

Analysis: Repeated guests cancel less. They are likely more familiar with hotel policies, have higher trust in the property, and may be traveling for similar reasons as before — reducing the chance of cancellation.

- Therefore, **assumption matches insights**

```
> cancel_repeated <- hotel_data %>%
+   group_by(is_repeated_guest) %>%
+   summarise(cancellation_rate = mean(is_canceled, na.rm = TRUE))
> cancel_repeated
# A tibble: 2 × 2
  is_repeated_guest cancellation_rate
              <int>             <dbl>
1                 0             0.378
2                 1             0.145
> 
```

STEP

2. Booking Changes by Repeated Guest Status

Assumption: Repeated guests will make fewer booking modifications due to more confidence and familiarity with the hotel.

- **Insight**:

- New Guests (0): 0.220
- Repeated Guests (1): 0.265

Analysis: Contrary to expectations, repeated guests make more booking changes. Repeated guests may feel more comfortable modifying their plans due to familiarity with the hotel's flexibility or may be booking for complex stays (e.g., extending visits, adding services), leading to more changes.

- Therefore, **assumption is different from insights**

```
> changes_repeated <- hotel_data %>%
+   group_by(is_repeated_guest) %>%
+   summarise(avg_changes = mean(booking_changes, na.rm = TRUE))
> changes_repeated
# A tibble: 2 × 2
  is_repeated_guest avg_changes
              <int>       <dbl>
1                 0       0.220
2                 1       0.265
```

STEP

3. Cancellation Rate by Waitlist Status

Assumption: Waitlisted guests are less certain and may cancel more often if not confirmed promptly.

- **Insight**:

- Not Waitlisted: 36.2%
- Waitlisted: 63.8%

Analysis: Guests on waitlists are almost twice as likely to cancel — showing clear risk in holding bookings on a waitlist without quick confirmations. Guests who are placed on a waitlist may explore alternative options while waiting, or may lose interest if confirmation takes too long, resulting in higher cancellations.

- Therefore, **assumption matches insights.**

```
> cancel_summary <- hotel_data %>%
+   group_by(waitlist_status) %>%
+   summarise(value = mean(is_canceled))
> cancel_summary
# A tibble: 2 × 2
  waitlist_status value
  <chr>           <dbl>
1 Not Waitlisted  0.362
2 Waitlisted      0.638
>
```

STEP

4. Booking Changes by Waitlist Status

Assumption: Waitlisted guests may make fewer changes, either due to uncertainty or preference to cancel outright.

- **Insight**:

- Not Waitlisted: 0.223
- Waitlisted: 0.159

Analysis: Booking modifications are fewer among waitlisted guests — they're either waiting or deciding to cancel altogether rather than updating details. Waitlisted guests may cancel instead of changing bookings. They might not make many changes because they're uncertain about the reservation's confirmation.

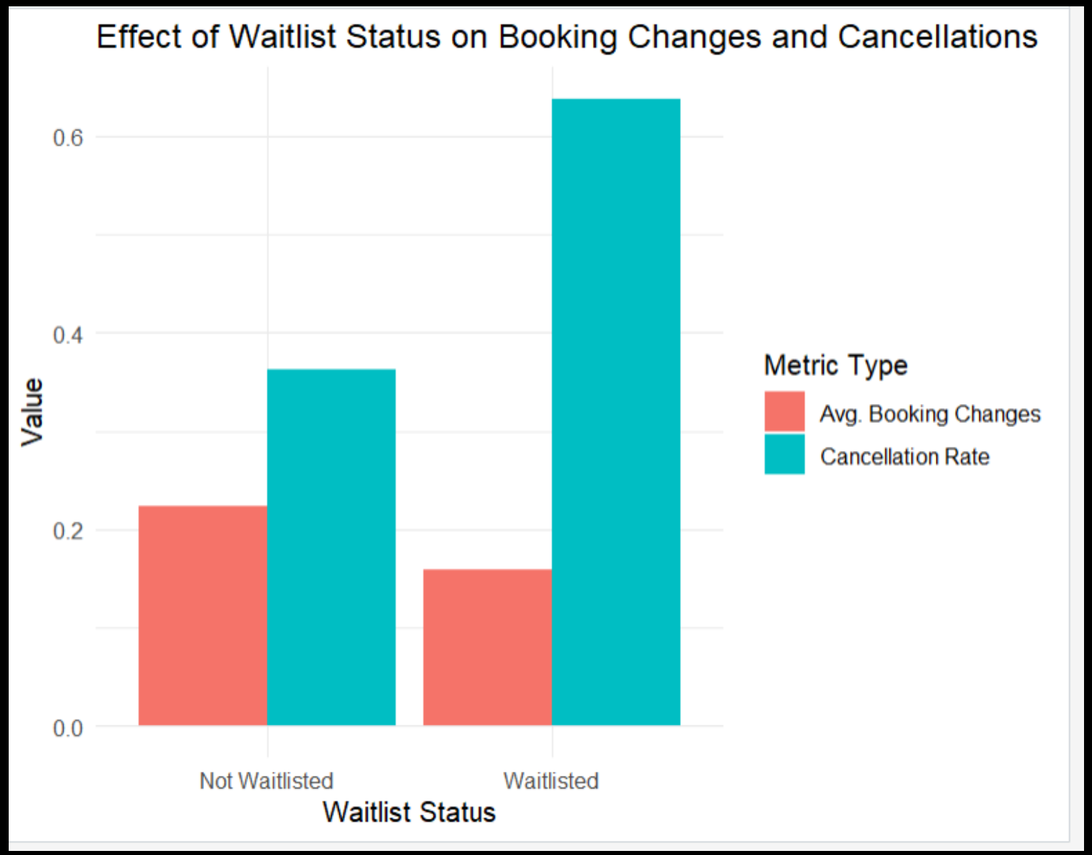
- Therefore, **assumption matches insights.**

```
> change_summary <- hotel_data %>%
+   group_by(waitlist_status) %>%
+   summarise(value = mean(booking_changes, na.rm = TRUE))
> change_summary
# A tibble: 2 × 2
  waitlist_status value
  <chr>           <dbl>
1 Not Waitlisted  0.223
2 Waitlisted      0.159
>
```

STEP

```
> combined_summary <- bind_rows(cancel_summary, change_summary)
> combined_summary
# A tibble: 4 × 3
  waitlist_status value metric
  <chr>           <dbl> <chr>
1 Not Waitlisted  0.362 Cancellation Rate
2 Waitlisted      0.638 Cancellation Rate
3 Not Waitlisted  0.223 Avg. Booking Changes
4 Waitlisted      0.159 Avg. Booking Changes
```

Combining all the outcomes for both metrics into a single plot



These insights offer valuable cues for hospitality teams:

- Build loyalty to reduce cancellations
- Engage waitlisted guests proactively to minimize loss
- Tailor communication for repeat guests who may modify bookings more but still stay loyal

CONCLUSION

1. Special Requests & Assigned Room Types → Satisfaction & Stay Duration
   - Matching reserved and assigned room types awith longer stays, suggesting higher satisfaction.
   - More special requests indicate longer or better-planned stays — possibly reflecting higher guest expectations or needs.

2. Lead Time & Cancellations → Revenue Impact
   - Longer lead times are associated with higher cancellation rates - affecting revenue predictability.
   - Guests who cancel often do so closer to the stay date in some cases, limiting the hotel's ability to rebook and impacting occupancy rates.

3. Country of Origin → Booking Behavior
   - International guests (e.g., Germany, UK) book much earlier and change bookings more often.
   - Portuguese guests (locals) cancel more frequently and modify bookings less — possibly due to flexible, last-minute planning.
   - Country-wise trends reflect cultural and travel behavior patterns that can inform marketing and forecasting.

4. Meal Packages → Modifications & Cancellations
   - Full Board guests cancel the most, going against the assumption that meal plans ensure commitment — possibly due to group or bulk bookings.
   - Guests without meals tend to make fewer changes, suggesting short or straightforward stays.
   - "Breakfast Only" is the most popular, likely due to balance of value and flexibility.

5. Waitlist & Repeated Guest Status → Booking Behavior
   - Repeated guests cancel less but make slightly more modifications — likely due to comfort and familiarity with hotel systems.
   - Waitlisted guests cancel much more and change bookings less, reflecting booking uncertainty and disengagement.
   - Loyalty (repeated stays) is a strong predictor of reliable behavior, while waitlisting requires better engagement strategies


FINANCIAL OUTLOOK
INCOME OVERVIEW
50
40
30
20
10
March
YOY PROFIT
67%
BUSINESS GROWTH
73%
WEBSITE TRAFFIC
INCOME EXPENSES
Total
$137,000
Expenses
$15,048

THANK YOU!