Code

HOTEL_BOOKING

eric

2/26/2020

Visit my github account

1 HOTEL BOOKING DATASET

2 INTRODUCTION

1. ***Ever wondered where and when is the best time to visit hotels***
2. ***This data analysis is very helpful in facilitating making of informed decisions.***

2.1 PROJECT OBJECTIVE :

To predict the best time,place and how to decision-make while there.

3 PREPARATION OF DATA & PACKAGES

3.1 Loading the necessary library packages

Code

```
library(dbplyr)
```

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.3.0
--
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts ------------------------------------------------ tidyverse_conflicts()
--
## x dplyr::filter() masks stats::filter()
## x dplyr::ident()  masks dbplyr::ident()
## x dplyr::lag()    masks stats::lag()
## x dplyr::sql()    masks dbplyr::sql()
```

Code

```
library(ggplot2)
library(gganimate)
library(plotrix)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
## The following object is masked from 'package:base':
##
##     date
```

Code

```
library(hrbrthemes)
```

```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these theme
s.
##       Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed
and
##       if Arial Narrow is not on your system, please see http://bit.ly/arialnarrow
```

Code

```
library(plotly)
```

```
##
```

```
## Attaching package: 'plotly'
## The following object is masked from 'package:ggplot2':
##
##     last_plot
## The following object is masked from 'package:stats':
##
##     filter
## The following object is masked from 'package:graphics':
##
##     layout
```

Code

```
library(dygraphs)
library(png)
library(grid)
library(bootstrap)
library(gapminder)
```

3.2 *Importing data from the csv dataset*

Code

```
HOTEL_BOOKING <- read.csv("E:/Documents/class/FINAL_PROJECT/hotel_bookings.csv",string
sAsFactors = F)
```

4 DATA CLEANING

4.1 changing month name into numeric.

Code

```
HOTEL_BOOKING$arrival_date_month <-match(HOTEL_BOOKING$arrival_date_month,month.name)
str(HOTEL_BOOKING$arrival_date_month)
```

```
##  int [1:119390] 7 7 7 7 7 7 7 7 7 7 ...
```

Code

```
table(HOTEL_BOOKING$arrival_date_month)
```

```
##
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12
##  5929  8068  9794 11089 11791 10939 12661 13877 10508 11160  6794  6780
```

Code

```
?match
```

```
## starting httpd help server ... done
```

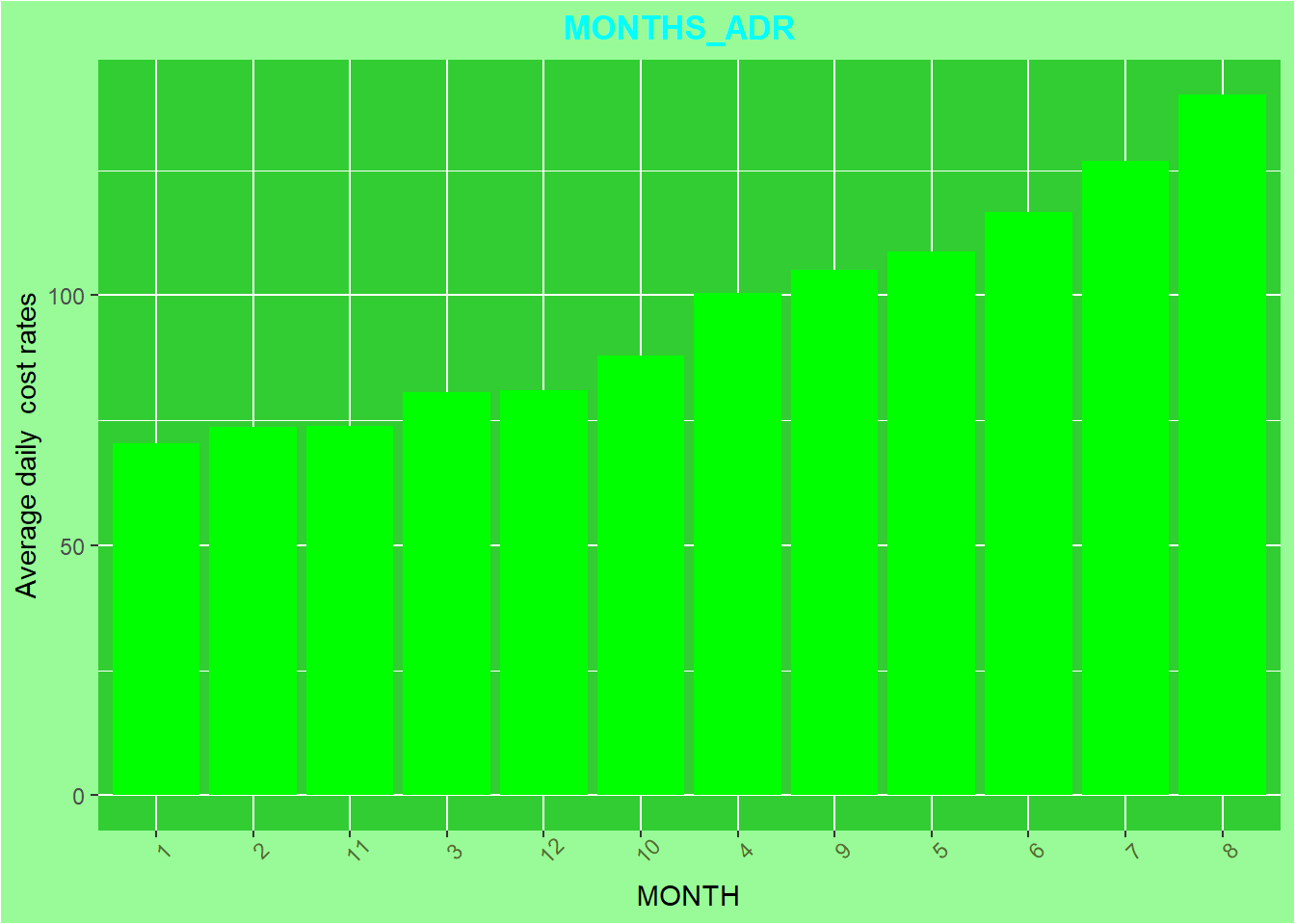
5 DATA ANALYSIS & VISUALISATION

5.1 MONTHLY ADR

- **ADR = AVERAGE DAILY RATE**
- Its clear that in months 6,7,8 average costing rates are very high.
- Fortunetely months 1,2,11,3,12 have the best rates while rates increase respectively.

Code

```
MONTHS_ADR <- HOTEL_BOOKING%>%
  select(adr,arrival_date_month)%>%
  group_by(arrival_date_month)%>%
  summarise(adr = mean(adr))%>%
  arrange(desc(adr))


#VISUALIZATION
ggplot(MONTHS_ADR,aes(reorder(x=as.factor(arrival_date_month),adr),y=adr))+
  geom_bar(stat = "identity",position="dodge",fill="green")+
  labs(title = "MONTHS_ADR  ",x="MONTH",y="Average daily  cost rates")+
  theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
  theme(panel.background = element_rect(fill = "limegreen"))+
  theme(axis.text.x = element_text(angle = 45,colour = "darkolivegreen"))+
  theme(plot.background = element_rect(fill="palegreen"))
```



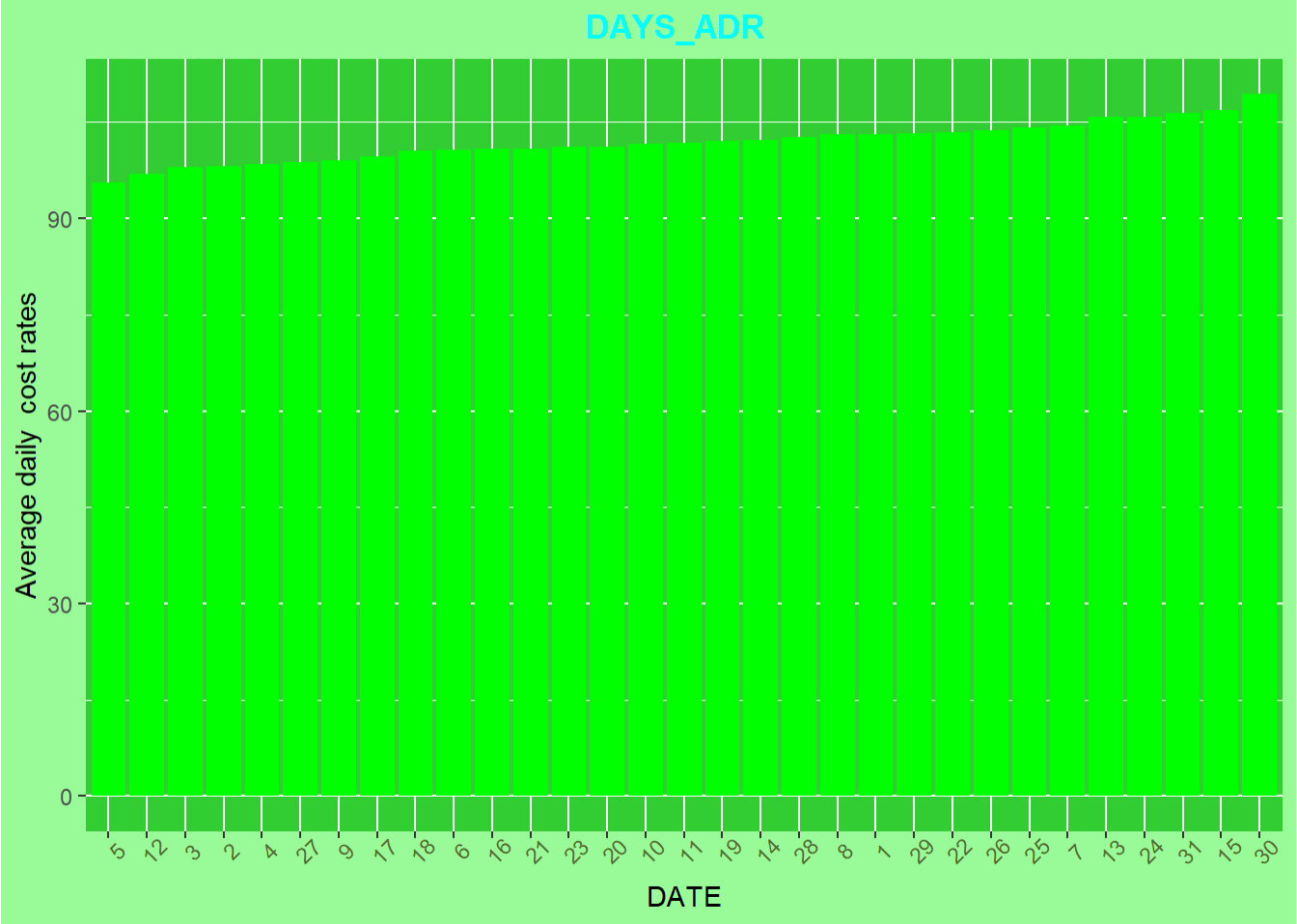
5.2 DAILY ADR

- Its show that towards the end of the month hotel costing rates increase

Code

```
DAYS_ADR <- HOTEL_BOOKING%>%
  select(adr,arrival_date_day_of_month)%>%
  group_by(arrival_date_day_of_month)%>%
  summarise(adr = mean(adr))%>%
  arrange(desc(adr))



#visualization
ggplot(DAYS_ADR,aes(reorder(x=as.factor(arrival_date_day_of_month),adr),y=adr))+
  geom_bar(stat = "identity",position="dodge",fill="green")+
```

```
  labs(title = "DAYS_ADR  ",x="DATE",y="Average daily  cost rates")+
  theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
  theme(panel.background = element_rect(fill = "limegreen"))+
  theme(axis.text.x = element_text(angle = 45,colour = "darkolivegreen"))+
  theme(plot.background = element_rect(fill="palegreen"))
```



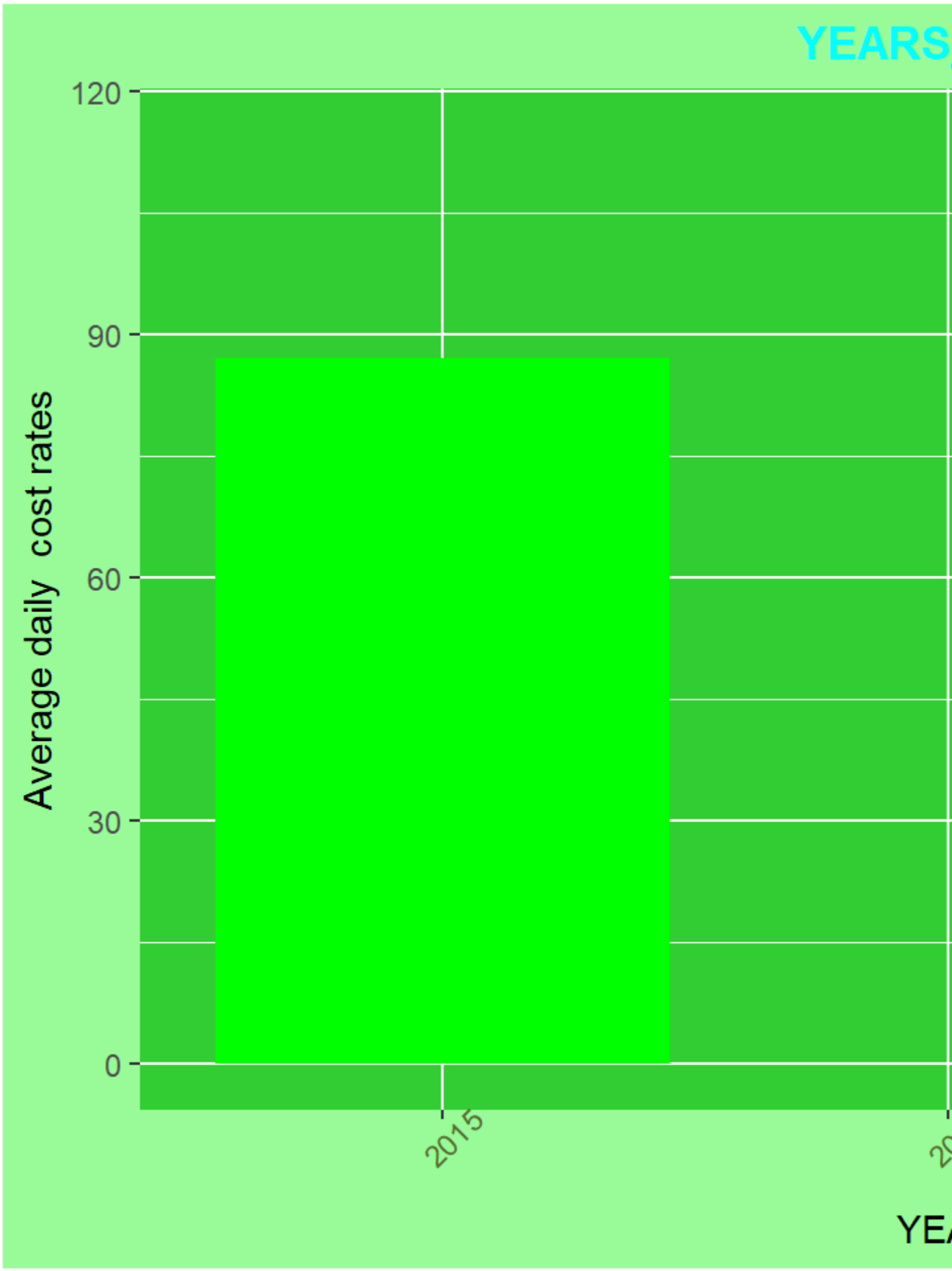
5.3 YEARLY ADR

- The above shows that through the years 2015,2016,2017 there has been a gradual increase in hotel costing rates.

Code

```
YEARS_ADR <- HOTEL_BOOKING%>%
  select(adr,arrival_date_year)%>%
  group_by(arrival_date_year)%>%
  summarise(adr = mean(adr))%>%
```

```
  arrange(desc(adr))


#VISUALIZATION
ggplot(YEARS_ADR,aes(reorder(x=as.factor(arrival_date_year),adr),y=adr))+
  geom_bar(stat = "identity",position="dodge",fill="green")+
  labs(title = "YEARS_ADR  ",x="YEAR",y="Average daily  cost rates")+
  theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
  theme(panel.background = element_rect(fill = "limegreen"))+
  theme(axis.text.x = element_text(angle = 45,colour = "darkolivegreen"))+
  theme(plot.background = element_rect(fill="palegreen"))+
  transition_states(arrival_date_year,transition_length = 2, state_length = 1 )+
  ease_aes('sine-in-out')
```


YEARS
120
90
60
30
0
Average daily cost rates
2015
YE

Code

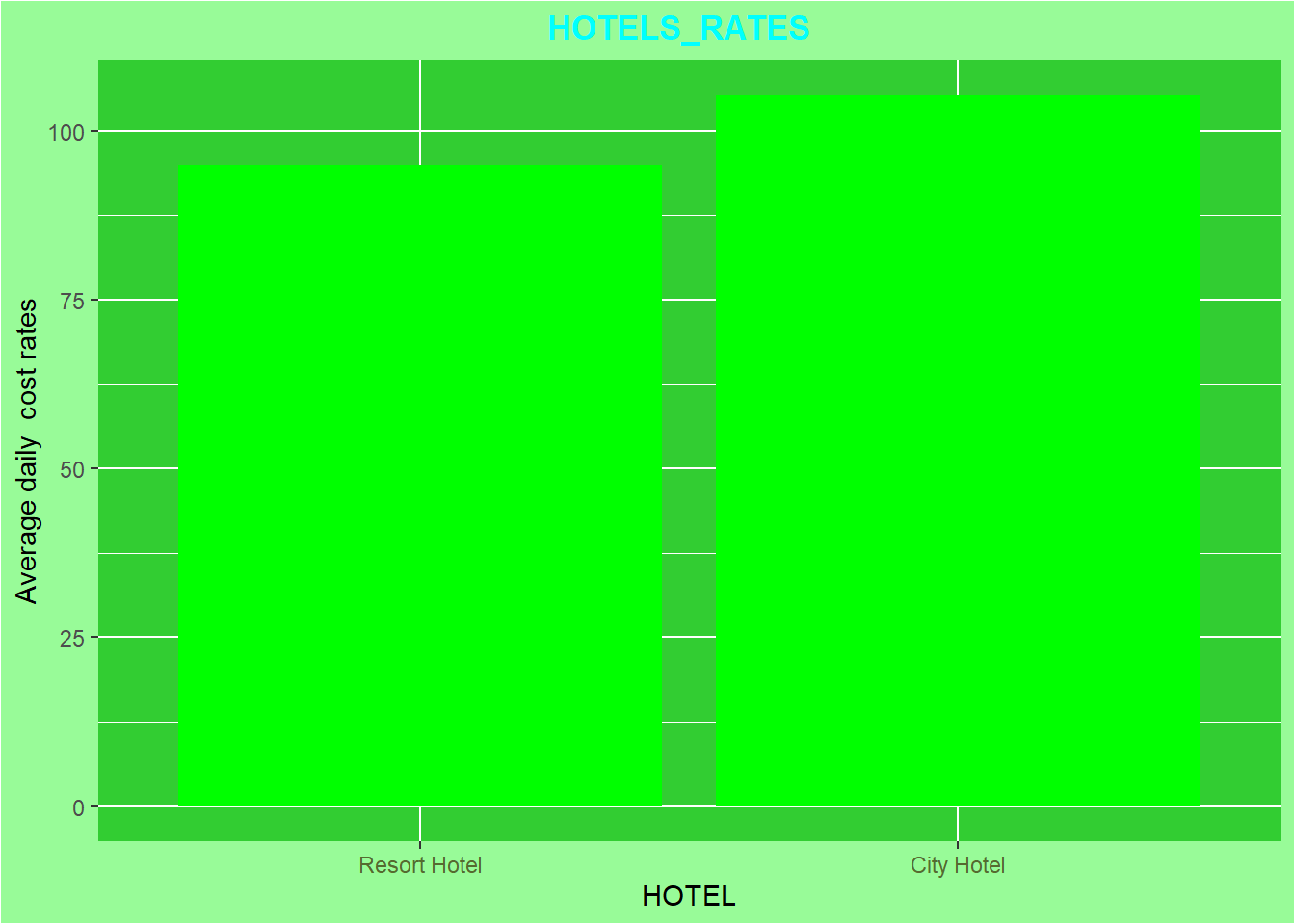
```
anim_save("years_adr_analysis.gif")
```

5.4 HOTELS ADR

- Its clear that CITY HOTEL has slightly a higher average costing RESORT HOTEL.

Code

```
HOTELS_RATES <- HOTEL_BOOKING%>%
  select(adr,hotel)%>%
  group_by(hotel)%>%
  summarise(adr = mean(adr))%>%
  arrange(desc(adr))



#VISUALIZATION


ggplot(HOTELS_RATES,aes(reorder(x=as.factor(hotel),adr),y=adr))+
  geom_bar(stat = "identity",position="dodge",fill="green")+
  labs(title = "HOTELS_RATES  ",x="HOTEL",y="Average daily  cost rates")+
  theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
  theme(panel.background = element_rect(fill = "limegreen"))+
  theme(axis.text.x = element_text(angle = 0,colour = "darkolivegreen"))+
  theme(plot.background = element_rect(fill="palegreen"))
```



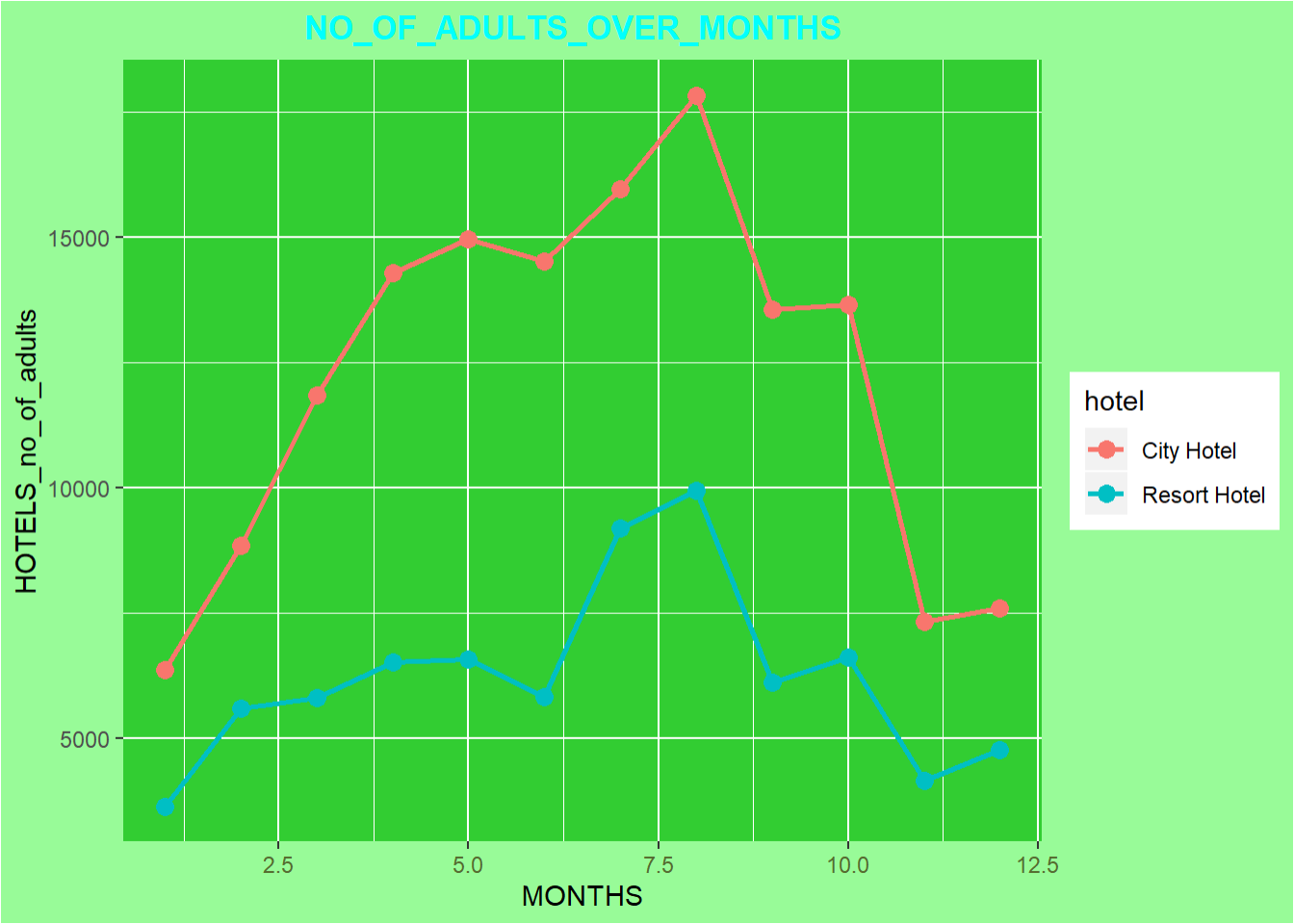
5.5 ADULTS VISITS OVER MONTHS

- It turns out that CITY HOTEL has relatively higher number of adults than RESORT HOTEL.
- However in all hotels the number of adults is higher in months 6,7,8 , while lower in months 1,2,3,11,12.

Code

```
NO_OF_ADULTS_OVER_MONTHS <- HOTEL_BOOKING%>%
  select(arrival_date_month,adults,hotel)%>%
  group_by(arrival_date_month,hotel)%>%
  summarise(no_of_adults = sum(adults))%>%
  arrange(desc(no_of_adults))


#VISUALIZATION
ggplot(NO_OF_ADULTS_OVER_MONTHS,aes(x=arrival_date_month,y=no_of_adults,col=hotel))+
geom_line(size=1)+
```

```
  geom_point(size=3)+
  geom_point(size=3,stat="identity",fill="green")+
  labs(title = "NO_OF_ADULTS_OVER_MONTHS  ",x="MONTHS",y="HOTELS_no_of_adults")+
  theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
  theme(panel.background = element_rect(fill = "limegreen"))+
  theme(axis.text.x = element_text(angle = 0,colour = "darkolivegreen"))+
  theme(plot.background = element_rect(fill="palegreen"))
```



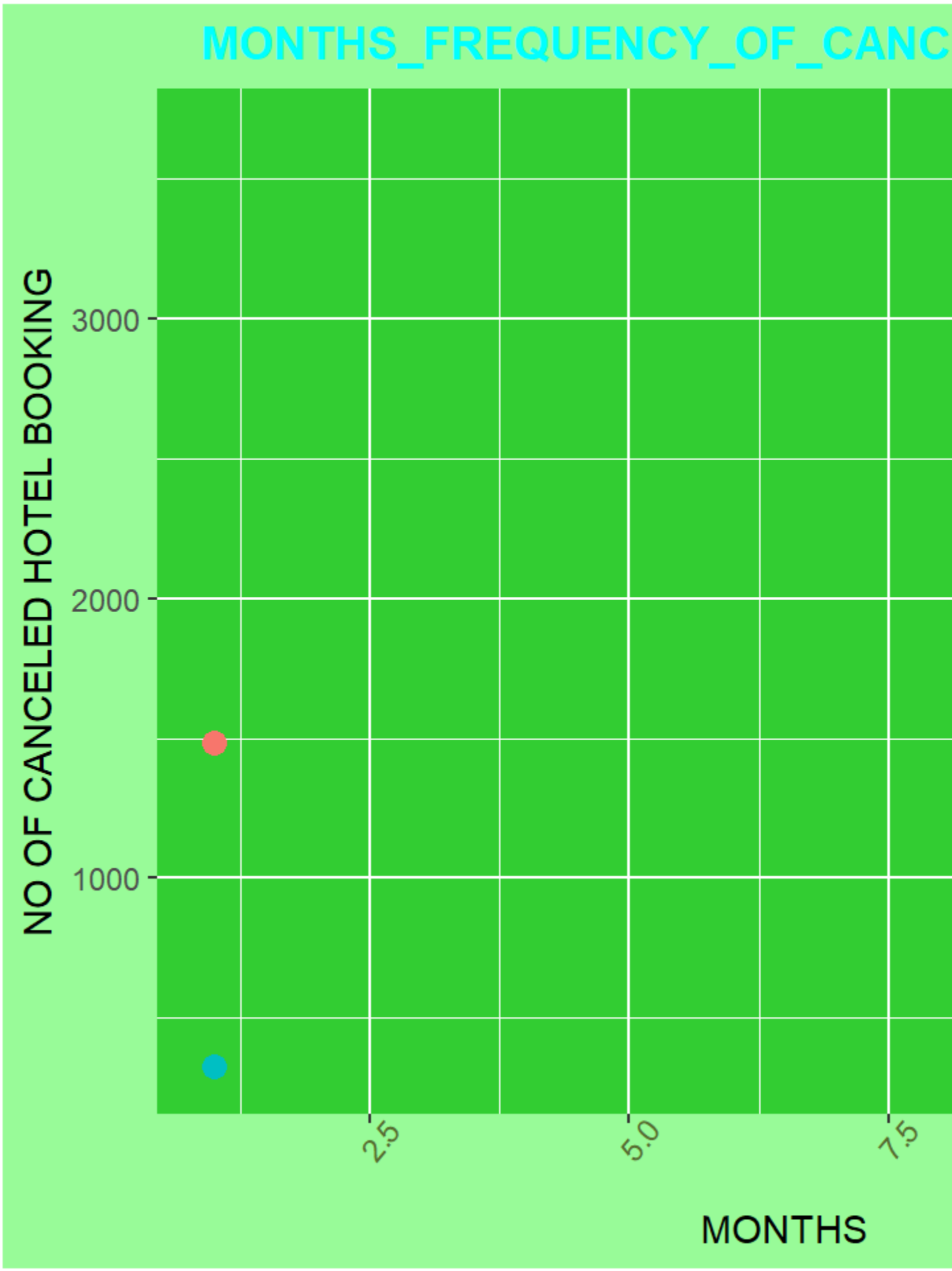
5.6 CANCELED_BOOKINGS_IN_HOTELS analysis

- Its observed that CITY HOTEL has a relatively higher number of cancelled hotel bookings than RESORT HOTEL.
- Precisely in months 5,6,7,8 booking cancelations are high in both hotels.

Code

```
MONTHS_FREQUENCY_OF_CANCELED_BOOKINGS_IN_HOTELS <- HOTEL_BOOKING%>%
  select(arrival_date_month,is_canceled,hotel)%>%
```

```
   filter(is_canceled==1)%>%
  group_by(arrival_date_month,hotel)%>%
  summarise(canceled_booking= n())%>%
  arrange(desc(canceled_booking))




 ggplot(MONTHS_FREQUENCY_OF_CANCELED_BOOKINGS_IN_HOTELS,aes(x=arrival_date_month,y=can
celed_booking ,color=hotel))+
   geom_line(size=1)+
   geom_point(size=3)+
   labs(title = "MONTHS_FREQUENCY_OF_CANCELED_BOOKINGS  ",x="MONTHS",y="NO OF CANCELED
HOTEL BOOKING")+
   theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
   theme(panel.background = element_rect(fill = "limegreen"))+
   theme(axis.text.x = element_text(angle = 50,colour = "darkolivegreen"))+
   theme(plot.background = element_rect(fill="palegreen"))+
   transition_reveal(arrival_date_month)
```


MONTHS_FREQUENCY_OF_CANC
NO OF CANCELED HOTEL BOOKING
3000
2000
1000
2.5
5.0
7.5
MONTHS

Code

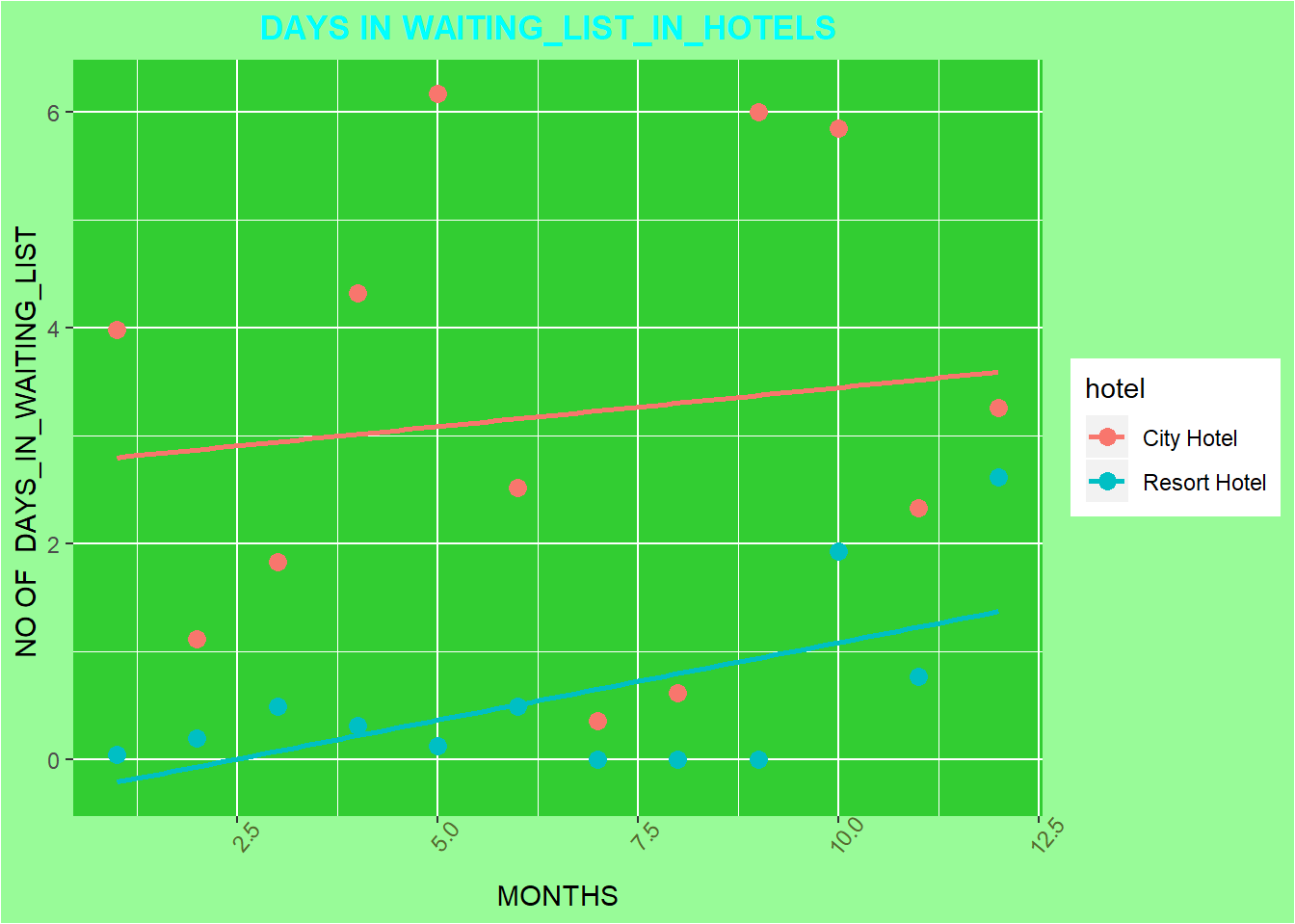
```
anim_save("no_of_canceled_bookings.gif")
```

5.7 WAIT DAYS

- Its observed that in CITY HOTEL a customer has to wait relatively longer for a hotel booking to be confirmed compared to RESORT HOTEL.

Code

```
DAYS_IN_WAITING_LIST_IN_HOTELS <- HOTEL_BOOKING%>%
   select(arrival_date_month,days_in_waiting_list,hotel)%>%
   group_by(arrival_date_month,hotel)%>%
   summarise(days_in_waiting_list=mean(days_in_waiting_list))%>%
   arrange(desc(days_in_waiting_list))



 #VISUALIZATION
 ggplot(DAYS_IN_WAITING_LIST_IN_HOTELS,aes(x=arrival_date_month,y=days_in_waiting_list
,color=hotel))+
   geom_point(size=3)+
   geom_smooth(method = "lm",se=F)+
   labs(title = "DAYS IN WAITING_LIST_IN_HOTELS  ",x="MONTHS",y="NO OF  DAYS_IN_WAITIN
G_LIST ")+
   theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
   theme(panel.background = element_rect(fill = "limegreen"))+
   theme(axis.text.x = element_text(angle = 50,colour = "darkolivegreen"))+
   theme(plot.background = element_rect(fill="palegreen"))
```



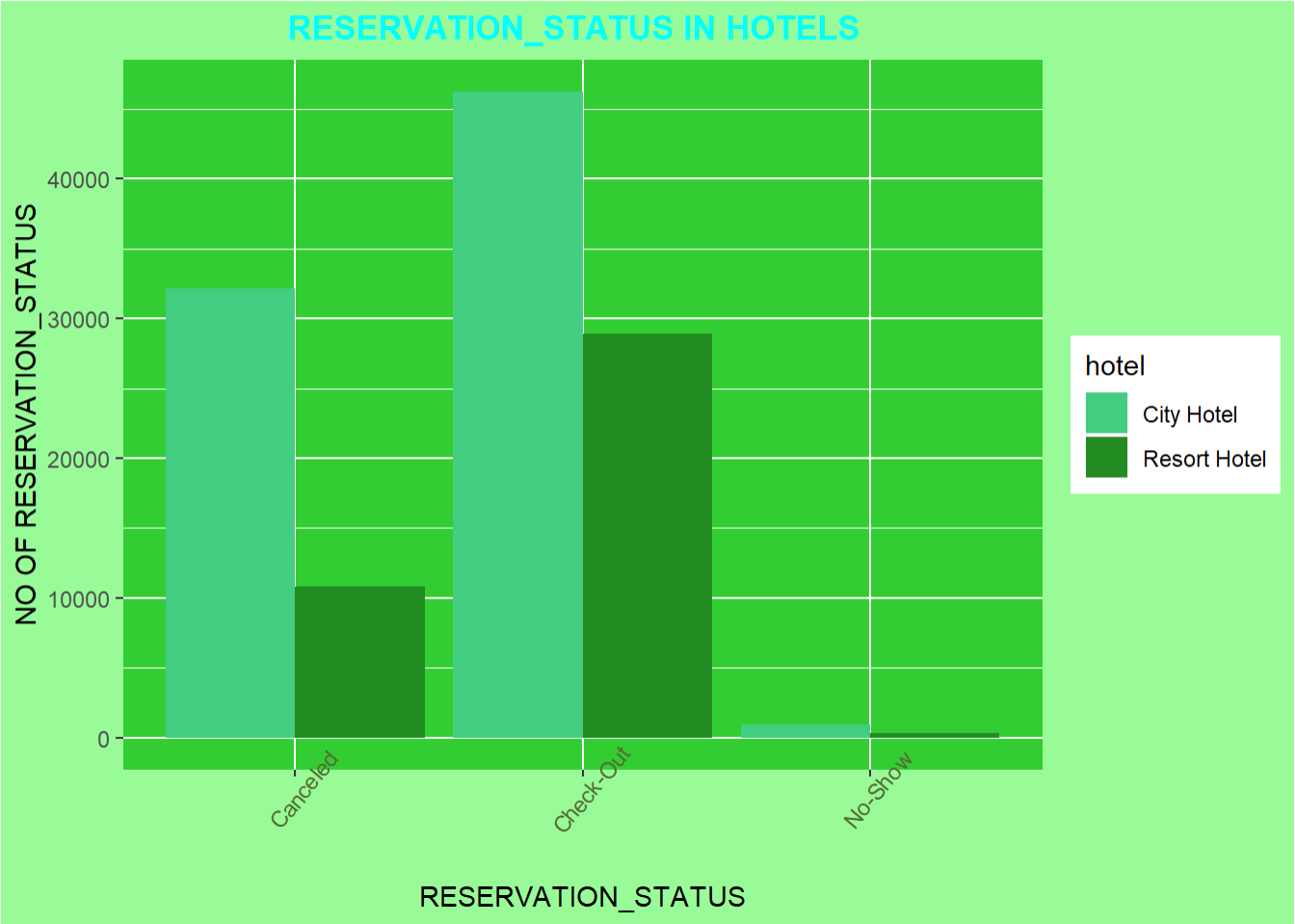
5.8 HOTEL RESERVATION STATUS

- The CITY HOTEL leads in canceled , check-outs and no-show hotel reservation statuses.
- This arguably because of its large scale service performance.

Code

```
RESERVATION_STATUS <- HOTEL_BOOKING%>%
   select(reservation_status,arrival_date_month,hotel)%>%
   group_by(reservation_status,hotel)%>%
   summarise(count= n())%>%
   arrange(desc(count))


 #VISUALIZATION
 ggplot(RESERVATION_STATUS,aes(y=count,x=reservation_status ,fill=hotel))+
   geom_bar(position="dodge",stat = "identity" )+
   scale_fill_manual(values = c("seagreen3","forestgreen"))+
```

```
   labs(title = "RESERVATION_STATUS IN HOTELS  ",x="RESERVATION_STATUS",y="NO OF RESER
VATION_STATUS")+
   theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
   theme(panel.background = element_rect(fill = "limegreen"))+
   theme(axis.text.x = element_text(angle = 50,colour = "darkolivegreen"))+
   theme(plot.background = element_rect(fill="palegreen"))
```



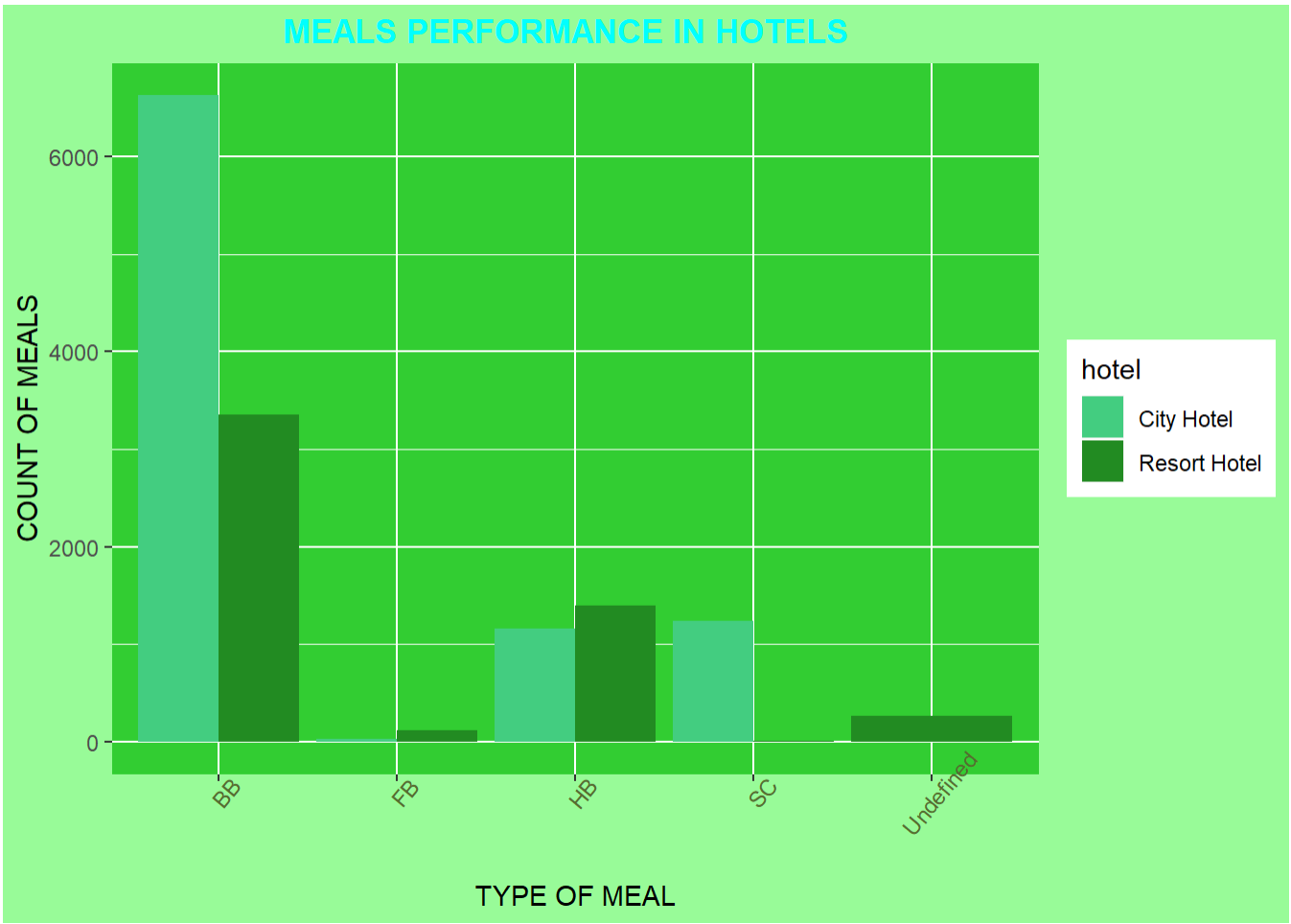
5.9 MEALS_perfomance

- From the above its clear that BB meal has the highest raqnking in both hotels.
- FB meal performs lowest in RESORT HOTEL.
- Undefined meal is lowest (not available) in city HOTEL.
- Generaly the counts of all meals are relatively higher from months 3 to 8.

Code

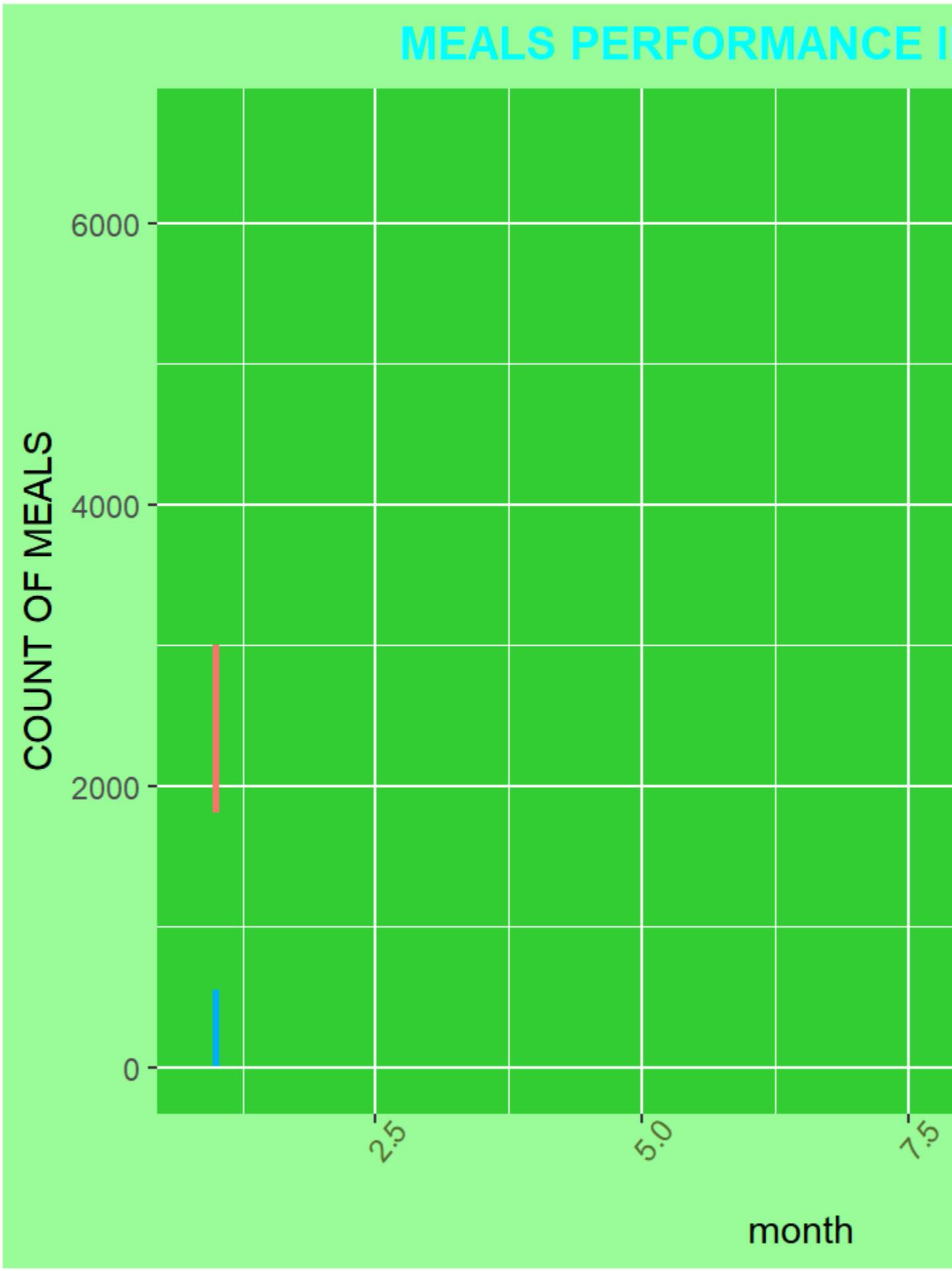
```
MEALS_perfomance <- HOTEL_BOOKING%>%
   select(meal,hotel,arrival_date_month)%>%
```

```
  group_by(meal,hotel,arrival_date_month)%>%
  summarise(count= n())%>%
  arrange(desc(count))



#VISUALIZATION
ggplot(MEALS_perfomance,aes(y=count,x=meal ,fill=hotel))+
  geom_bar(position="dodge",stat = "identity" )+
  scale_fill_manual(values = c("seagreen3","forestgreen"))+
  labs(title = "MEALS PERFORMANCE IN HOTELS  ",x="TYPE OF MEAL",y="COUNT OF MEALS")+
  theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
  theme(panel.background = element_rect(fill = "limegreen"))+
  theme(axis.text.x = element_text(angle = 50,colour = "darkolivegreen"))+
  theme(plot.background = element_rect(fill="palegreen"))
```



Code

```
 ggplot(MEALS_perfomance,aes(y=count,x=arrival_date_month ,col=meal,group=meal))+
   geom_line(size=1,stat = "identity" )+
   #scale_fill_manual(values = c("seagreen3","forestgreen","cyan","darkgreen","paleturquoise"))+
   labs(title = "MEALS PERFORMANCE IN HOTELS  ",x="month",y="COUNT OF MEALS")+
   theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
   theme(panel.background = element_rect(fill = "limegreen"))+
   theme(axis.text.x = element_text(angle = 50,colour = "darkolivegreen"))+
   theme(plot.background = element_rect(fill="palegreen"))+
   transition_reveal(arrival_date_month)
```


MEALS PERFORMANCE I
COUNT OF MEALS
6000
4000
2000
0
2.5
5.0
7.5
month

Code

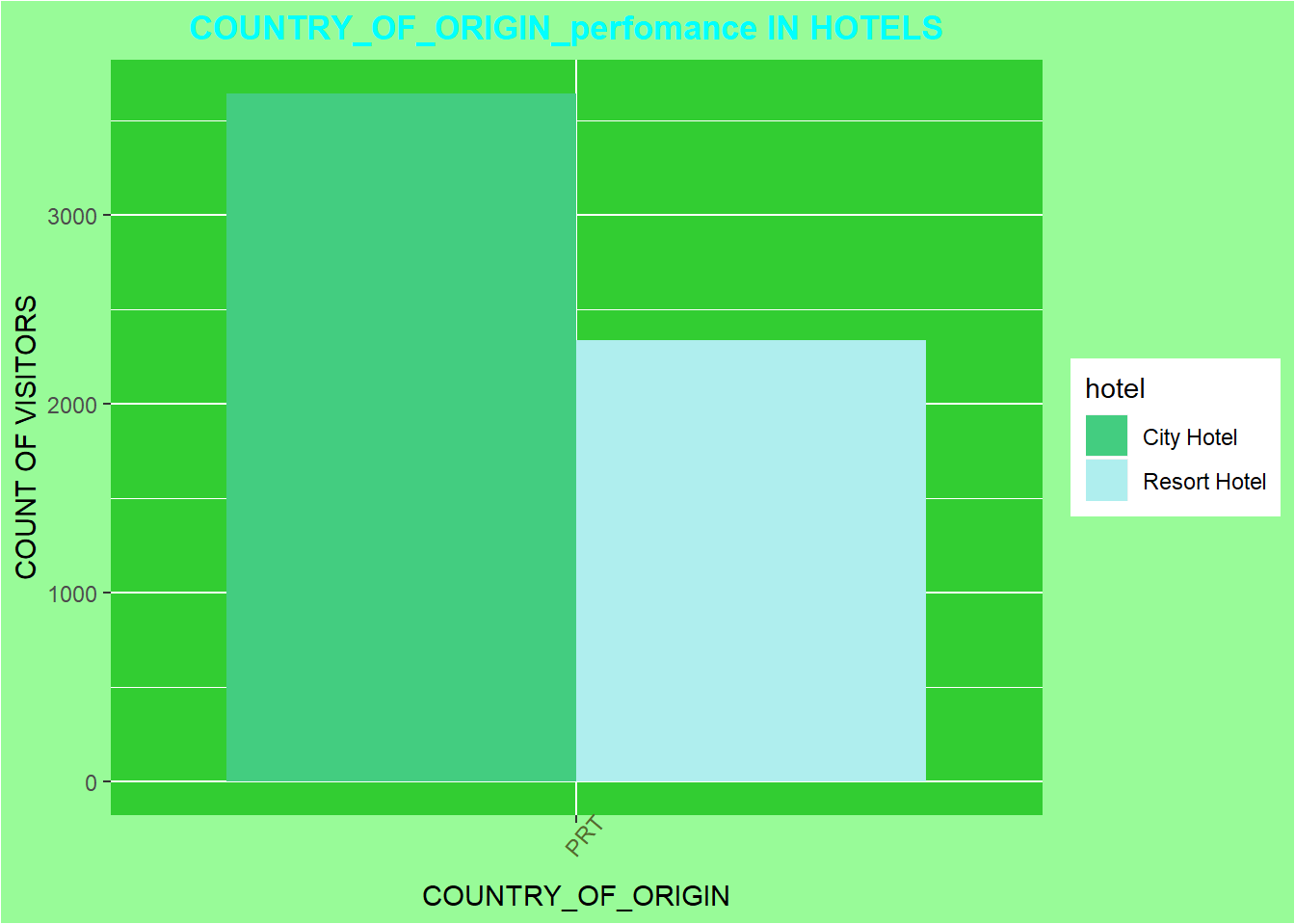
```
  anim_save("meals_performance.gif")
```

5.10 COUNTRY_OF_ORIGIN_perfomance

- The above analysis proves that out of the top 10 in rank countrys of origin PRT citizen are the highest numbers in visiting the two hotels however more in city hotel.
- Again out of the top 10 in rank countrys :DEU ,FRA and ITA citizens never pay visits to RESORT HOTEL .
- On the other hand citizens from country IRL never visits CITY HOTEL.

Code

```
COUNTRY_OF_ORIGIN_perfomance <- HOTEL_BOOKING%>%
   select(country,arrival_date_month,hotel)%>%
   group_by(country,arrival_date_month,hotel)%>%
   summarise(count= n())%>%
   arrange(desc(count))


 #DATA SORTING
 #Get a high performing sample of the data which is too huge.
 #Hence get the top 10 in rank.
 HEAD <- head(COUNTRY_OF_ORIGIN_perfomance,10)
 #get the bottom 10 in rank.
 TAIL <- tail(COUNTRY_OF_ORIGIN_perfomance,10)


 #VISUALIZATION


 ggplot(HEAD,aes(y=count,x=country ,fill=hotel))+
   geom_bar(position="dodge",stat = "identity" )+
   scale_fill_manual(values = c("seagreen3","paleturquoise"))+
   labs(title = "COUNTRY_OF_ORIGIN_perfomance IN HOTELS  ",x="COUNTRY_OF_ORIGIN",y="CO
UNT OF VISITORS")+
   theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
   theme(panel.background = element_rect(fill = "limegreen"))+
   theme(axis.text.x = element_text(angle = 50,colour = "darkolivegreen"))+
   theme(plot.background = element_rect(fill="palegreen"))
```



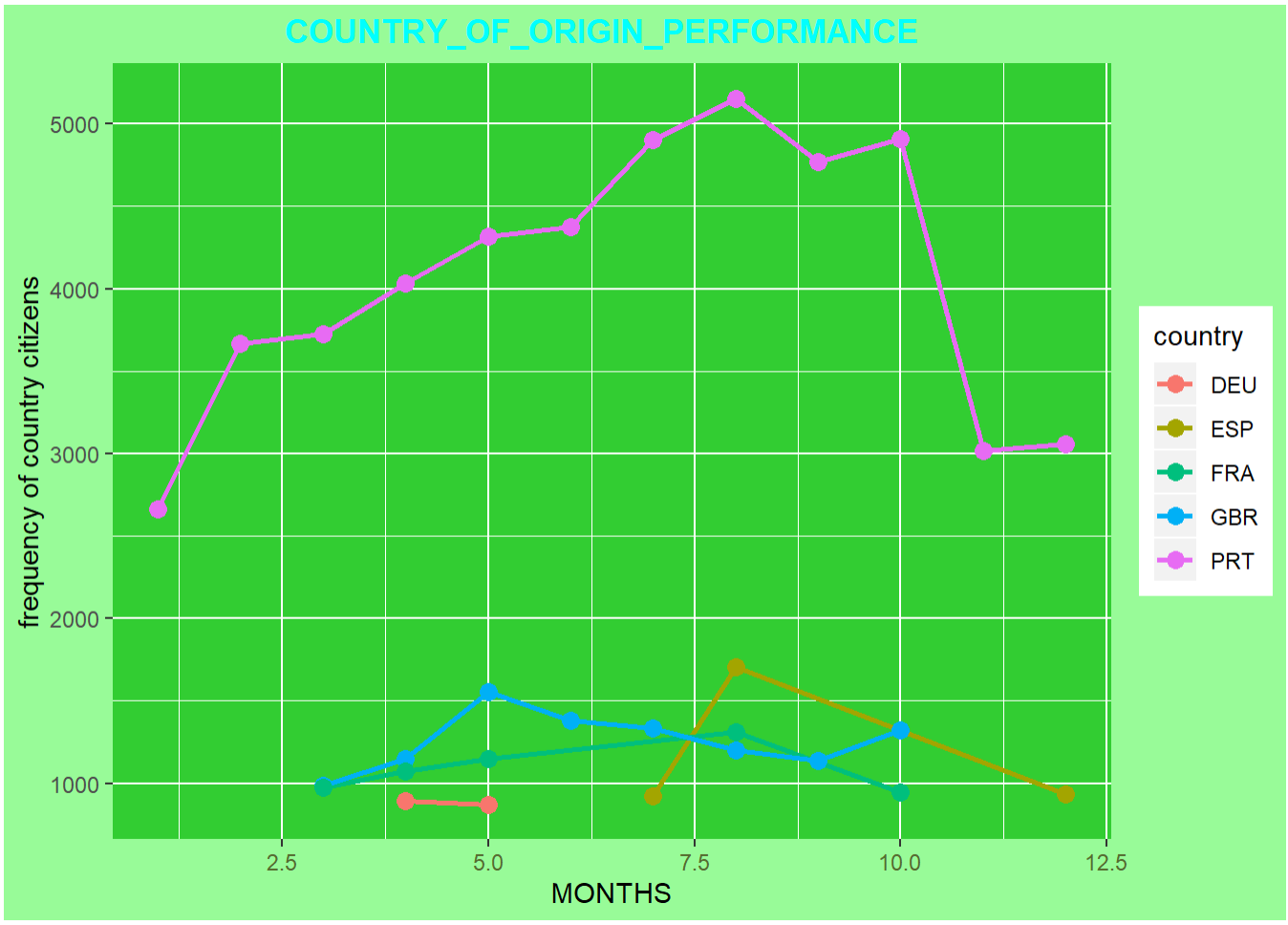
5.11 ORIGIN_PERFORMANCE_OVER_MONTHS

- FROM the above analysis it was noted that visitors from PRT were visting through the months but where in large numbers in month ; 7 ,8,9,10.
- However visitors from DEV only made visits in month 4.
- Visitors from ESP made most visits in months :6 ,7,8 .
- Visitors from GBR made most visits in month 5.

Code

```
COUNTRY_OF_ORIGIN_PERFORMANCE_OVER_MONTHS <- HOTEL_BOOKING%>%
   select(country,arrival_date_month,hotel)%>%
   group_by(country,arrival_date_month)%>%
   summarise(count= n())%>%
   arrange(desc(count))
 #Sampling of data to take the top 30 countrys of origin.
 HEAD2 <- head(COUNTRY_OF_ORIGIN_PERFORMANCE_OVER_MONTHS,30)
```

```
#VISUALIZATION
ggplot(HEAD2,aes(x=arrival_date_month,y=count,col=country))+
  geom_line(size=1)+
  geom_point(size=3,stat="identity")+
  labs(title = "COUNTRY_OF_ORIGIN_PERFORMANCE  ",x="MONTHS",y="frequency of country c
itizens")+
  theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
  theme(panel.background = element_rect(fill = "limegreen"))+
  theme(axis.text.x = element_text(angle = 0,colour = "darkolivegreen"))+
  theme(plot.background = element_rect(fill="palegreen"))
```

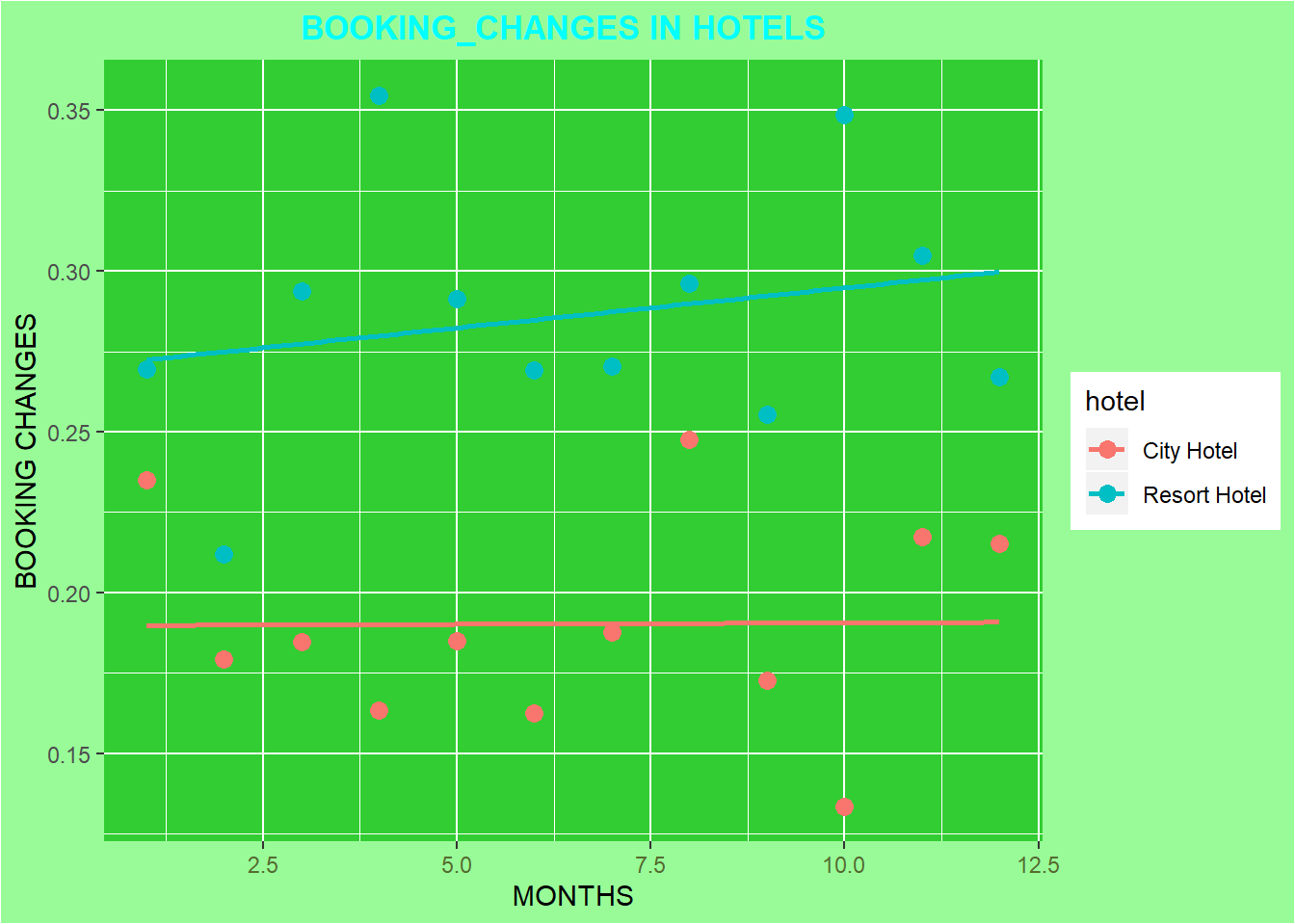


5.12 BOOKING_CHANGES ANALYSIS

- Its been proved that hotel booking changes are most likely to happen in RESORT HOTEL than in CITY hotel.

Code

```
BOOKING_CHANGES <- HOTEL_BOOKING%>%
   select(booking_changes,arrival_date_month,hotel)%>%
   group_by(arrival_date_month,hotel)%>%
   summarise(booking_changes=  mean(booking_changes))%>%
   arrange(desc(booking_changes))


 #VISUALIZATION
 ggplot(BOOKING_CHANGES,aes(x=arrival_date_month,y=booking_changes,col=hotel))+
   geom_smooth(method = "lm",se=F)+
   geom_point(size=3)+
   #scale_fill_manual(values = c("seagreen3","forestgreen"))+
   labs(title = "BOOKING_CHANGES IN HOTELS  ",x="MONTHS",y="BOOKING CHANGES")+
   theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
   theme(panel.background = element_rect(fill = "limegreen"))+
   theme(axis.text.x = element_text(angle = 0,colour = "darkolivegreen"))+
   theme(plot.background = element_rect(fill="palegreen"))
```



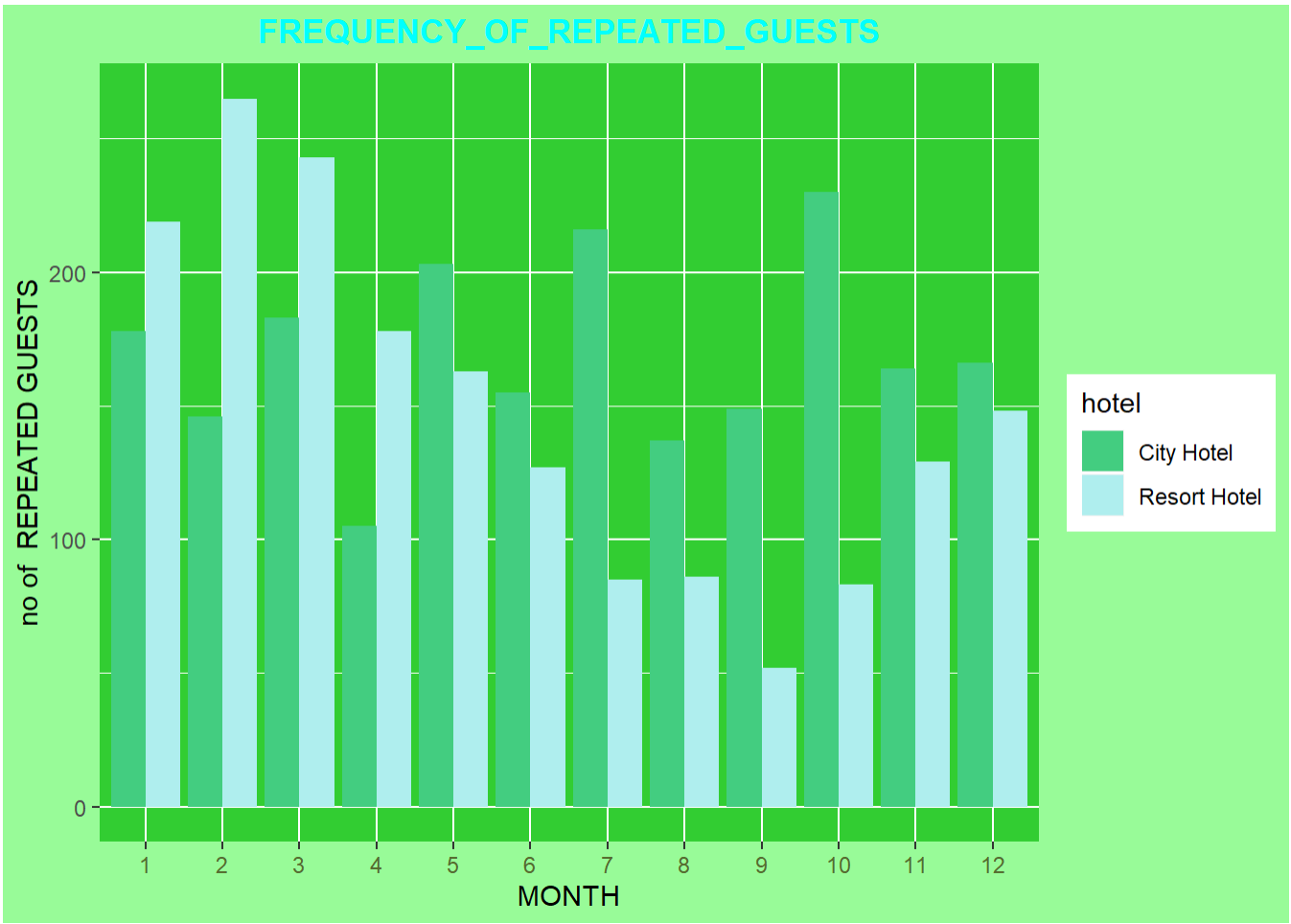
5.13 FREQUENCY_OF_REPEATED_GUESTS

- The above analysis proves that relatively many of the repeated visitors obted to visit both hotels in months :1,2 and 3.
- Again CITY HOTEL repeated guests obted to visit the hotel on months :7 and 10 .
- On the other hand RESORT HOTEL`S repeated guest obted to visit the hotel in months :1,2 and 3.

Code

```
FREQUENCY_OF_IS_REPEATED_GUESTS <- HOTEL_BOOKING%>%
   select(is_repeated_guest,hotel,arrival_date_month)%>%
   filter(is_repeated_guest==1)%>%
   group_by(arrival_date_month,hotel)%>%
   summarise(count= n())%>%
   arrange(desc(count))
```

```
 #VISUALIZATION


 ggplot(FREQUENCY_OF_IS_REPEATED_GUESTS,aes(x=as.factor(arrival_date_month)  ,y=count,
fill=hotel))+
   geom_bar(position="dodge",stat = "identity" )+
   scale_fill_manual(values = c("seagreen3","paleturquoise"))+
   labs(title = " FREQUENCY_OF_REPEATED_GUESTS ",x="MONTH",y="no of  REPEATED GUESTS")
+
   theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
   theme(panel.background = element_rect(fill = "limegreen"))+
   theme(axis.text.x = element_text(angle = 0,colour = "darkolivegreen"))+
   theme(plot.background = element_rect(fill="palegreen"))
```

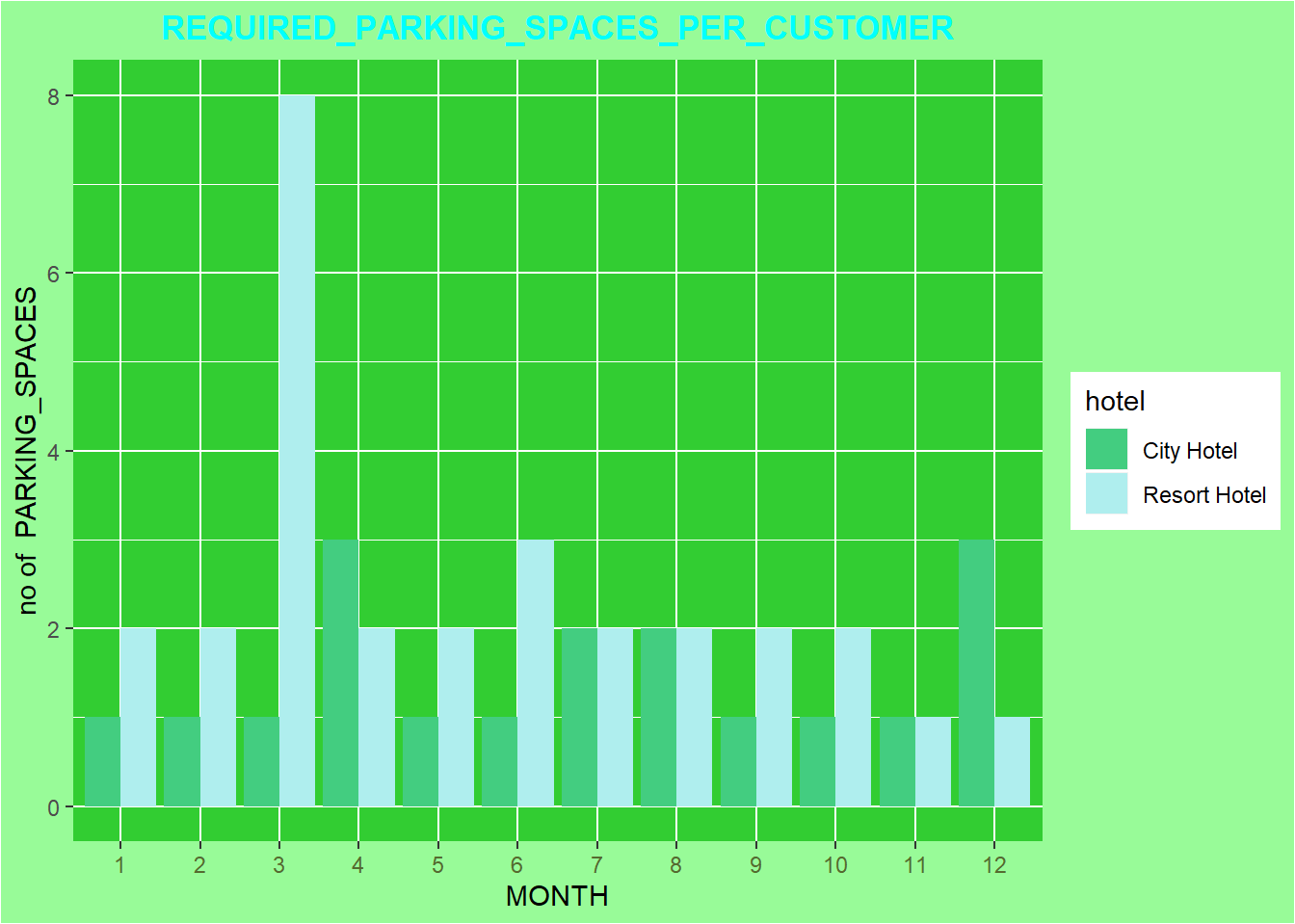


5.14 REQUIRED PARKING SPACES PER CUSTOMER

- It obsevered that people with many cars visits RESORT HOTEL on month :3.
- However visitors with many cars in CITY HOTEL tend to visit the hotel in month :4 and 12.
- In both hotels low car taffic is experienced in month :11 .

Code

```
REQUIRED_PARKING_SPACES_PER_CUSTOMER <- HOTEL_BOOKING%>%
   select(required_car_parking_spaces,arrival_date_month,hotel)%>%
   group_by(arrival_date_month,hotel)%>%
   arrange(desc(required_car_parking_spaces))




 #VISUALISATION


 ggplot(REQUIRED_PARKING_SPACES_PER_CUSTOMER ,aes(x=as.factor(arrival_date_month)  ,y=
required_car_parking_spaces,fill=hotel))+
   geom_bar(position="dodge",stat = "identity" )+
   scale_fill_manual(values = c("seagreen3","paleturquoise"))+
   labs(title = " REQUIRED_PARKING_SPACES_PER_CUSTOMER ",x="MONTH",y="no of  PARKING_S
PACES")+
   theme(plot.title = element_text(hjust = 0.5,colour = "cyan",face = "bold"))+
   theme(panel.background = element_rect(fill = "limegreen"))+
   theme(axis.text.x = element_text(angle = 0,colour = "darkolivegreen"))+
   theme(plot.background = element_rect(fill="palegreen"))
```



6 CONCLUSION

- From the above dataset analysis and information therein CITY HOTEL seems the best option.
- However months :1 and 2 are the most cost effective months.
- Other advantages of the CITY HOTEL are :large scale business operations which guarantee quality services ,and social adventure due to its high visitors diversity.

Visit my github account