Data-Driven Future for Sun Country

Team 18

10/20/2019

The Problem Given and Our Approach

Our Task

Our consulting team was contracted to provide Sun Country with non-obvious patterns to better understand the airlineâ€™s customers while addressing the challenges they articulate. Currently, Sun Country relies on anecdotes to differentiate between their customers instead of a practice of rigor. The challenge in this task is that there are multiple patterns available to observe. In order to best aid Sun Country in making informed decisions to better understand the customers, our consulting team provides a high-level overview of the general customer qualities amongst all Sun Country passengers and then focus on the analysis. Limiting our analysis to a specific level of success will allow us to provide Sun Country with more concise insights.

Current Situation

Based on our client, one major concern about Sun Country is the lack of customer knowledge which hinders their marketing and advertising decisions. Looking at the top booking channels within the existing data, we noticed that even though many customers are booking through the SCA website, Sun Country is facing strong competition from outside booking resources.

Thus, our team decided to dive deeper into the SCA website booking to see what current challenges are being faced by Sun Country. We divided the data into SCA website booking and Non-SCA website booking. The following three problems are identified from our analysis:

1. Volume and earnings do not align for the SCA website
2. SCA website experienced a dip in performance since February 2014.
3. Not All Ufly members book through the SCA website

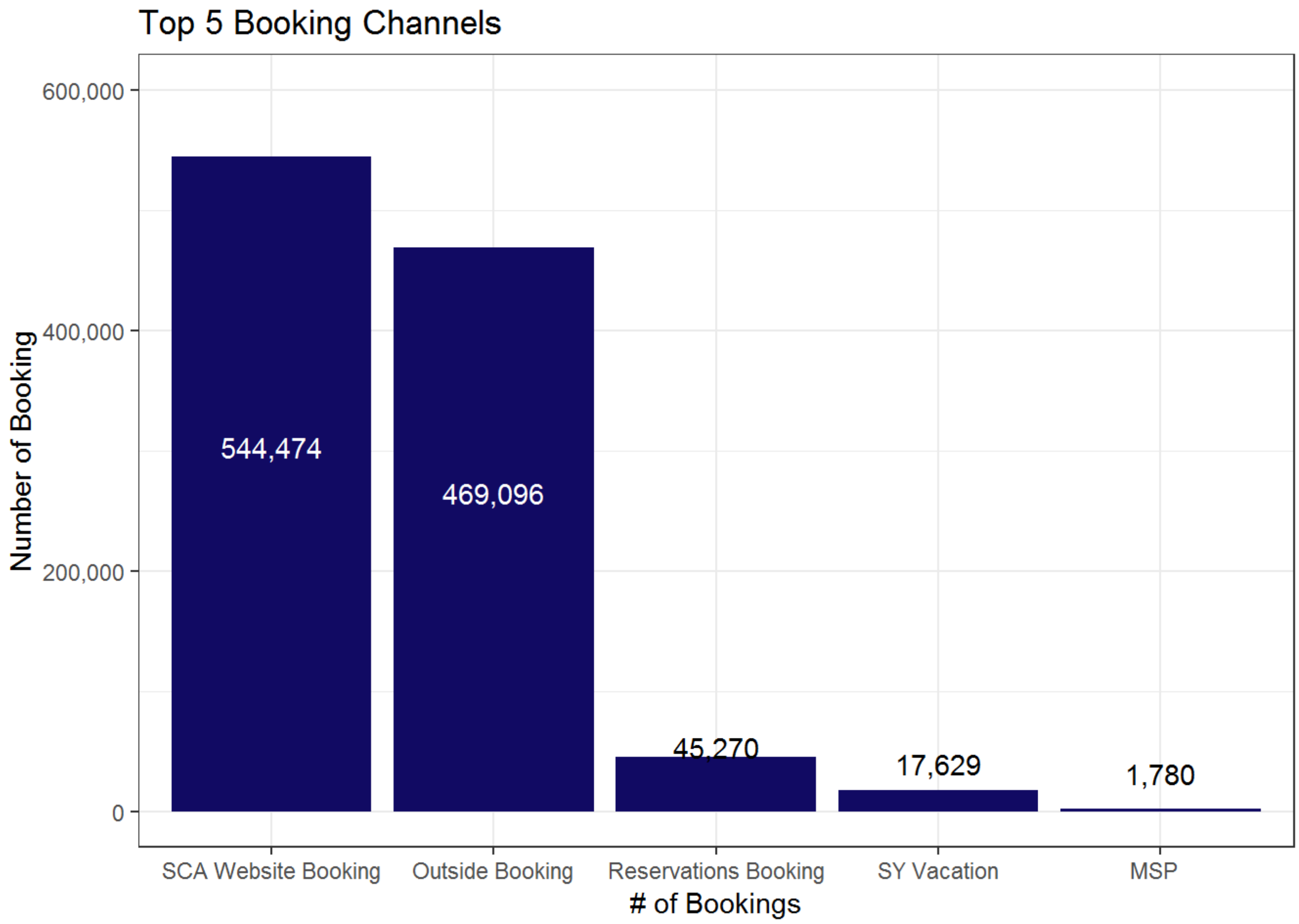
First, we looked at the total number of bookings and total revenue generated from the SCA website booking and Non-SCA website bookings.

```
library(magrittr)
library(RSQLite)
library(DBI)
library(tidyr)
library(tidyverse)
library(classInt)
library(Hmisc)
library(ggplot2)
library(dplyr)
library(sqldf)
```

```
df <- read.csv('SunCountry.csv')
sun_raw <- df
```

```
df1 <- sun_raw %>%
  filter(TotalDocAmt != 0) %>%
  mutate(flag = ifelse(BookingChannel == 'SCA Website Booking', 'SCA Website', 'Others')) %>%
  group_by(PNRLocatorID, BookingChannel, flag) %>%
  summarise(avg_doc_price = mean(TotalDocAmt))

# number of PNR by different channels
PNR_count <- df1 %>%
  group_by(BookingChannel) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  slice(1:5) %>%
  arrange(desc(count))
```

```
ggplot(PNR_count, aes(reorder(BookingChannel, -count), count)) +
  geom_bar(stat="identity", fill = '#110a63')  +
  scale_y_continuous(name="Number of Booking", limits = c(0, 600000), labels = scales::comma)
+
  ggtitle('Top 5 Booking Channels') +
  labs(x = "# of Bookings", y = "Channels") +
  geom_text(aes(label= scales::comma(count)), position = position_stack(vjust = 0.5), vjust=
 -1.20, col = c('white', 'white', 'black', 'black', 'black' )) +
  theme_bw()
```



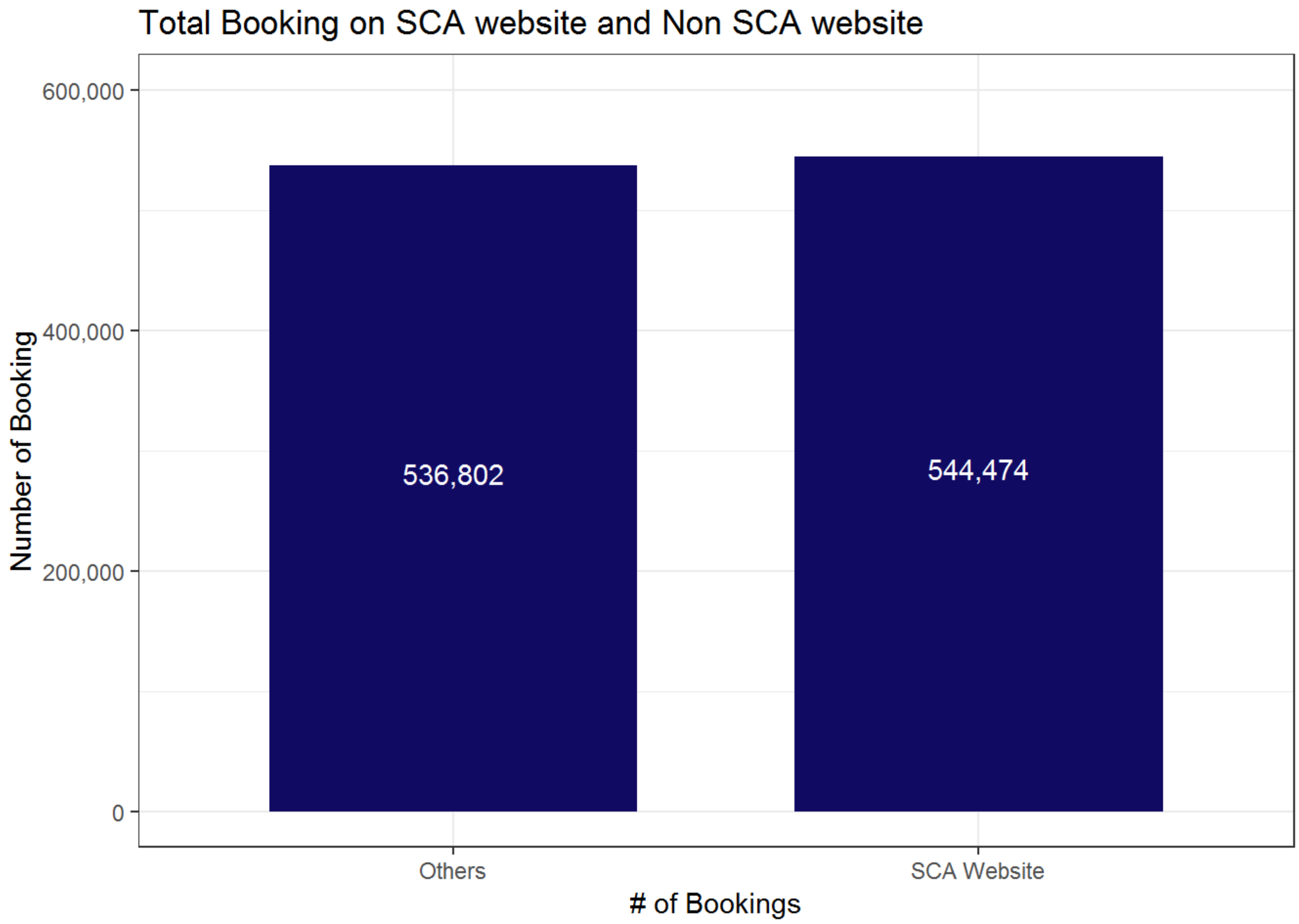
Thus, our team decided to dive deeper into the SCA website booking to see what current challenges are being faced by Sun Country. We divided the data into SCA website booking and Non-SCA website booking. The following three problems are identified from our analysis:

Volume and earnings do not align for the SCA website The SCA website experienced worse performance since February 2014. Not All Ufly members book through the SCA website

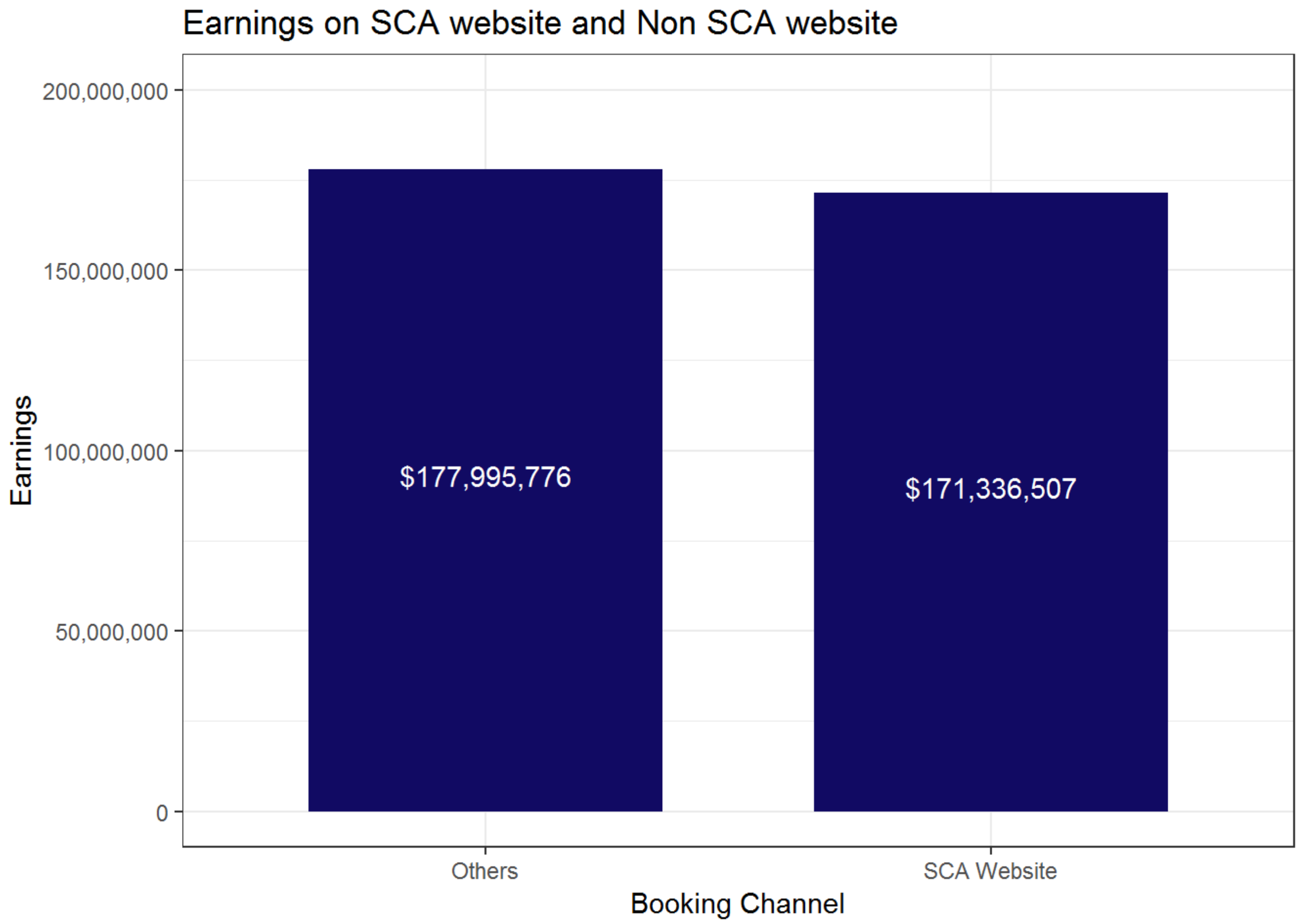
First, we looked at the total number of bookings and total revenue generated from the SCA website booking and Non-SCA website bookings.

```
SCA_count <- df1 %>%
  group_by(flag) %>%
  summarise(num = n())

SCA_total <- df1 %>%
  group_by(flag) %>%
  summarise(total_amt = round(sum(avg_doc_price)), 0)
```

```
ggplot(SCA_count, aes(flag, num)) +
  geom_bar(stat="identity", width = 0.7, fill = '#110a63') +
  scale_y_continuous(name="Number of Booking", limits = c(0, 600000), labels = scales::comma)
+
  geom_text(aes(label= scales::comma(num)), position = position_stack(vjust = 0.5), vjust= -
0.20, color = "white") +
  ggtitle('Total Booking on SCA website and Non SCA website') +
  labs(x = "# of Bookings", y = "Channels") +
  theme_bw()
```



```
ggplot(SCA_total, aes(flag, total_amt)) +
  geom_bar(stat="identity", width = 0.7, fill = '#110a63') +
  geom_text(aes(label= scales::dollar(total_amt)), position = position_stack(vjust = 0.5), vj
ust= -0.20, color = "white")+
  ggtitle('Earnings on SCA website and Non SCA website') +
  xlab('Booking Channel') + ylab('Revenue From Channels') +
  theme_bw() +
  scale_y_continuous(name="Earnings", limits = c(0, 200000000), labels = scales::comma)
```

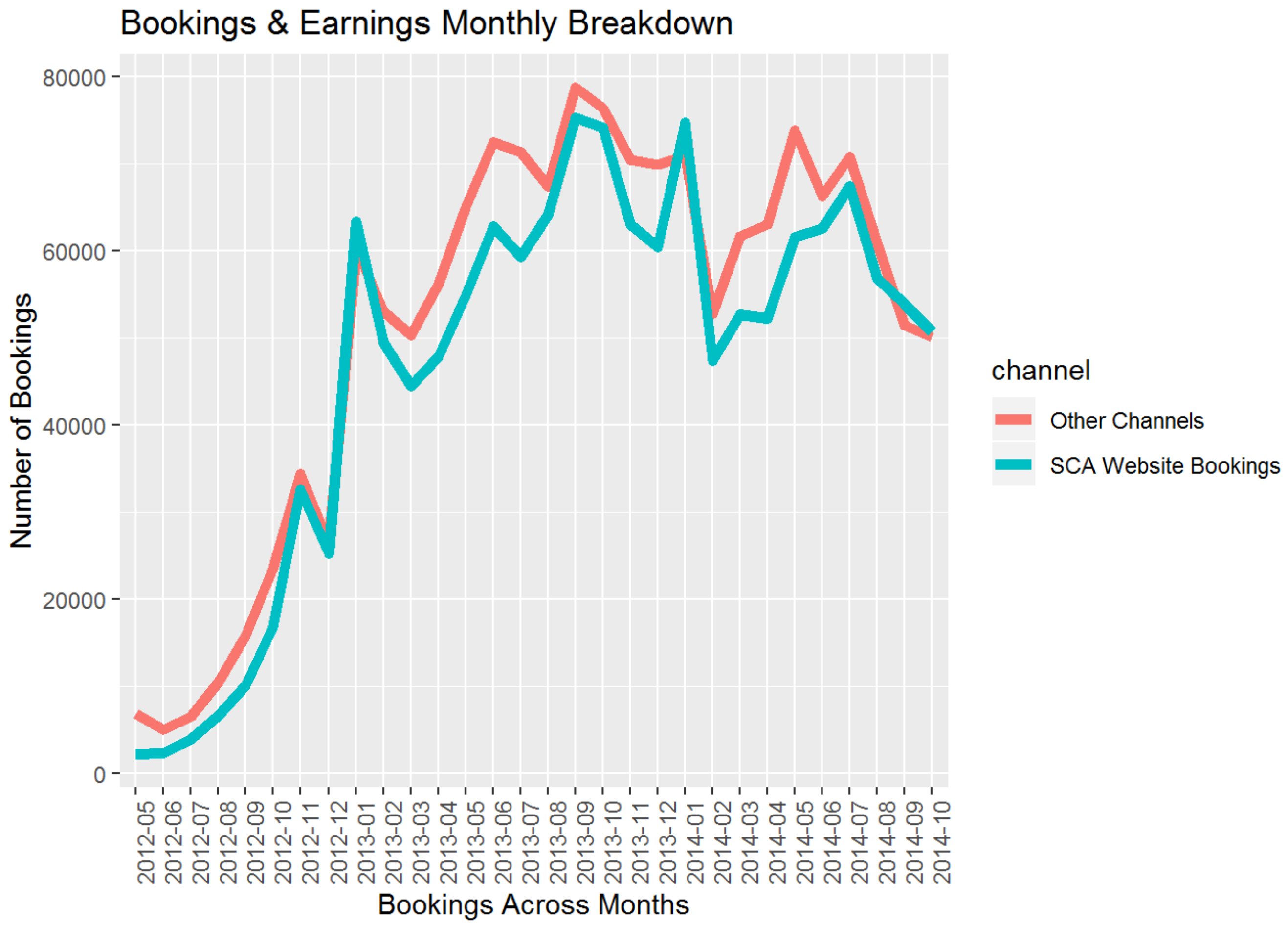


From these graphs, we observed our first point. We see that even though the SCA website has more booking volume than the non-SCA websites, the revenue generated from the SCA websites is less than that of the non-SCA websites. Next, we broke down the booking amount and revenue by month

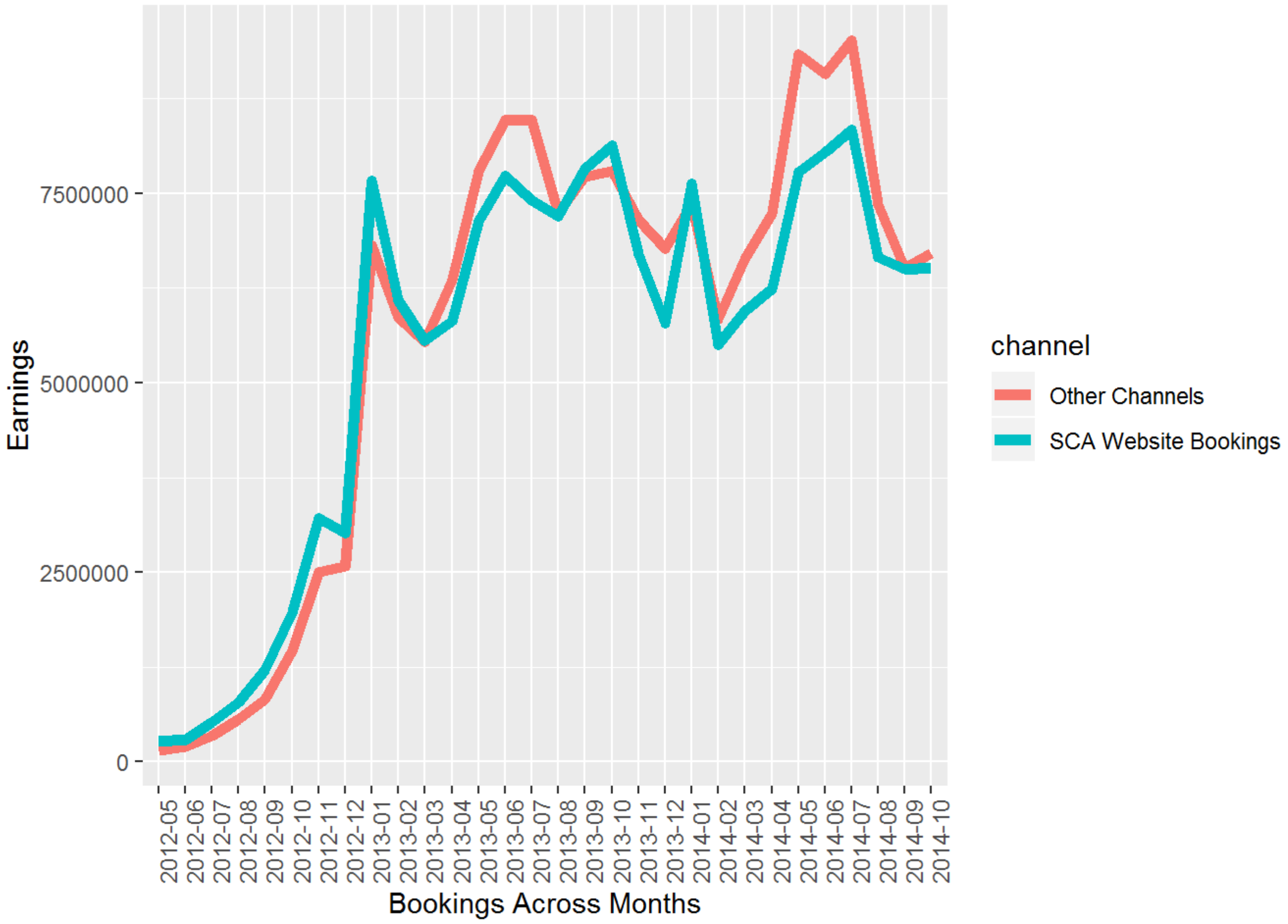
When looking at the booking volume and revenue per month plot, we observe our second point.

```
Price_Check <- sun_raw %>%
  filter(TotalDocAmt != 0) %>%
  group_by(PNRLocatorID, TotalDocAmt) %>%
  summarise(count = n()) %>%
  group_by(PNRLocatorID) %>%
  summarise(avg_doc_price = mean(TotalDocAmt))

channel <- sun_raw %>%
  mutate(booking_group = ifelse(BookingChannel == 'SCA Website Booking', 'SCA Website Bookings','Other Channels')) %>%
  filter(TotalDocAmt != 0) %>%
  group_by(PNRLocatorID) %>%
  summarise(booking_group_1 =min(booking_group))

count <- sun_raw %>%
  filter(TotalDocAmt != 0) %>%
  group_by(PNRLocatorID) %>%
  summarise(count_of_tickets = n())


month <- sqldf("select PNRLocatorID, min(PNRCreateDate) as booking_date from sun_raw where TotalDocAmt != 0 group by PNRLocatorID")
month$booking_month <- format(as.Date(month$booking_date, format = "%Y-%m-%d"),"%Y-%m")
month$booking_date <- NULL


table_1 <- cbind(month, price =Price_Check$avg_doc_price, channel= channel$booking_group_1, count_of_tickets = count$count_of_tickets)

booking_month_channel <- table_1 %>%
  group_by(booking_month, channel) %>%
  summarise(price = sum(price), count_of_tickets = sum(count_of_tickets))
```

```
ggplot(booking_month_channel %>%
         filter(booking_month >= '2012-05' &  booking_month <= '2014-10'),
       aes(x=booking_month, y=count_of_tickets, col=channel,group = factor(channel))) +
  geom_line(size=2) + theme(axis.text.x= element_text(angle=90)) + xlab('Bookings Across Months') +
  ylab('Number of Bookings') + ggtitle('Bookings & Earnings Monthly Breakdown')
```



```
ggplot(booking_month_channel %>% filter(booking_month >= '2012-05' &  booking_month <= '2014-
10'),
       aes(x=booking_month, y=price, col=channel,group = factor(channel))) +
  geom_line(size=2) + theme(axis.text.x= element_text(angle=90)) +
  xlab('Bookings Across Months') + ylab('Earnings')
```

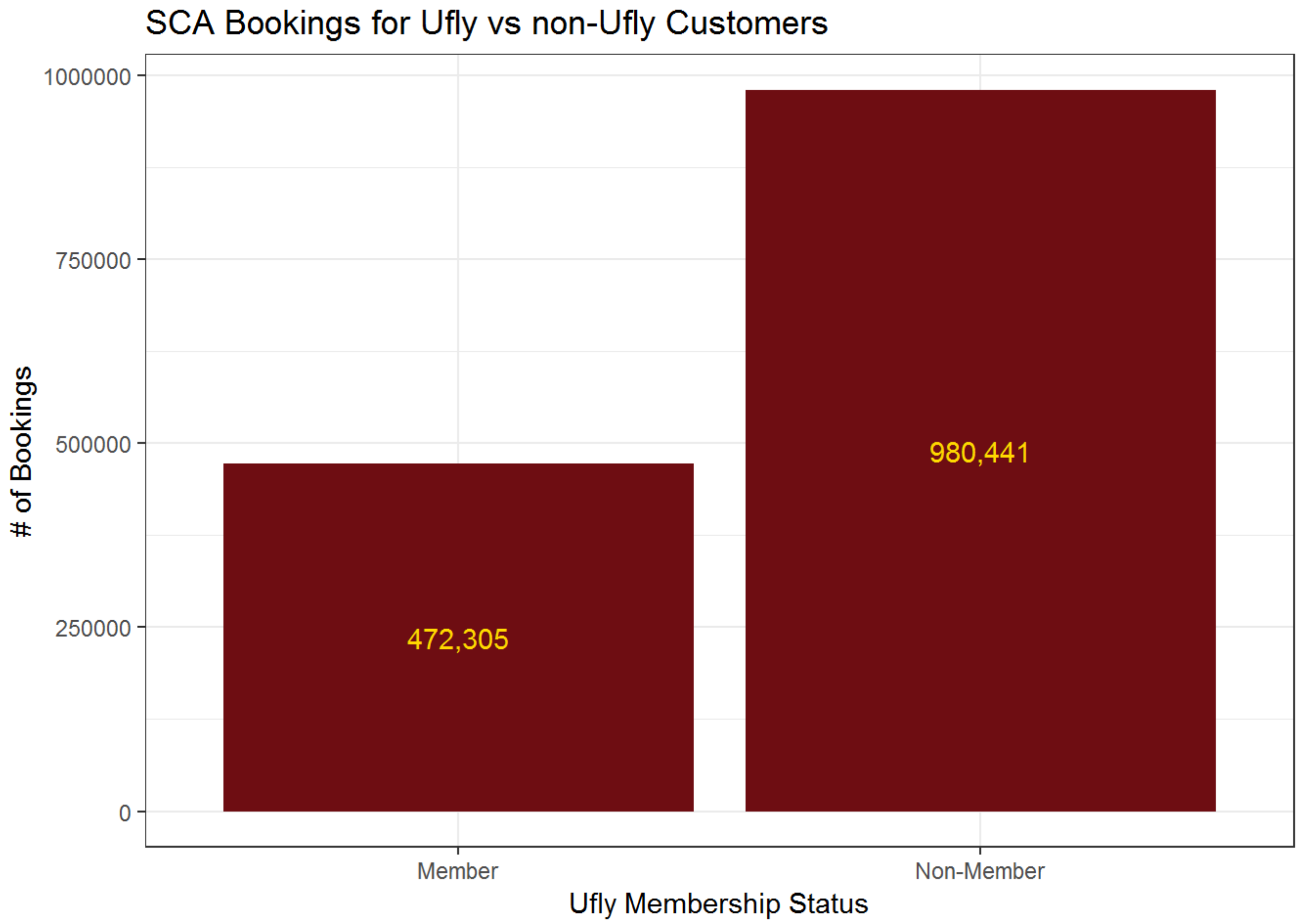


While the SCA websites experience fluctuations before February 2014, after February 2014 the SCA website is booking less than non-SCA websites and earning less revenue than non-SCA booking sites.

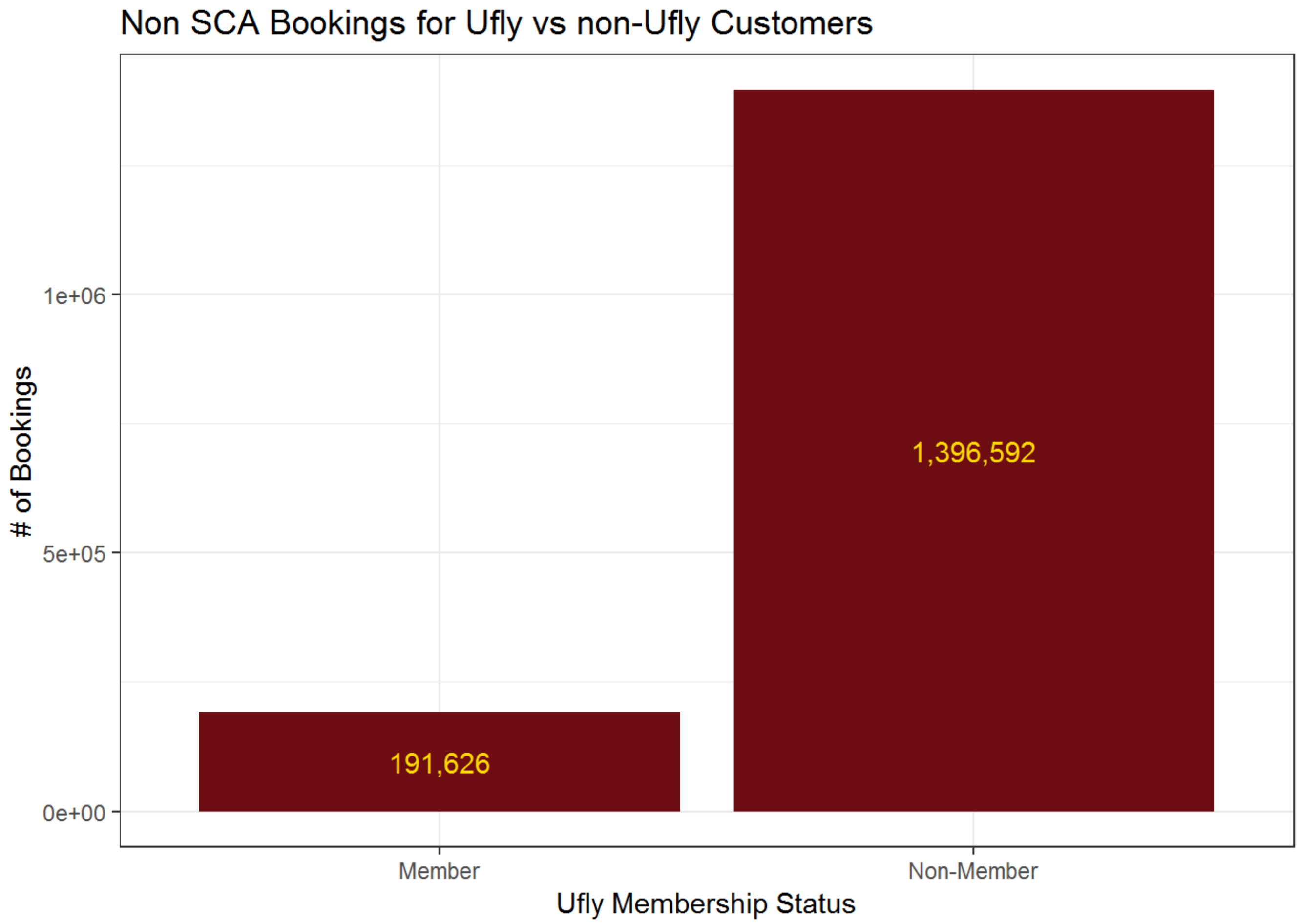
Additionally, we would like to see if our Ufly loyalty program is casting a positive effect by driving more people to book through the SCA website booking.

```
library(dplyr)
SCA <- sun_raw %>% filter(BookingChannel == 'SCA Website Booking' & TotalDocAmt != 0)

Price_Check <- sun_raw %>%
  filter(TotalDocAmt != 0) %>%
  group_by(PNRLocatorID, TotalDocAmt) %>%
  summarise(num = n()) %>%
  group_by(PNRLocatorID) %>%
  summarise(avg_doc_price = mean(TotalDocAmt))



SCA_UFly <- SCA %>%
  mutate(Ufly_Flag = ifelse(UflyMemberStatus == "Elite" | UflyMemberStatus == "Standard", "Me
mber" , "Non-Member"),
         Customers = paste(EncryptedName, birthdateid, sep="_")) %>%
  left_join(Price_Check, by = "PNRLocatorID") %>%
  group_by(Ufly_Flag) %>%
  summarise(num_customers = n_distinct(Customers),
            Num_PNR = n(),
            Revenue = sum(avg_doc_price))%>%
  mutate(Ticket_Per_Customer = round(Num_PNR/num_customers, 2),
         Rev_Per_Customer = round(Revenue/num_customers, 2),
         Rev_Per_PNR = round(Revenue/Num_PNR, 2))

non_SCA <- sun_raw %>%
  filter(BookingChannel != 'SCA Website Booking' & TotalDocAmt != 0)

non_SCA_UFly <- non_SCA %>%
  mutate(Ufly_Flag = ifelse(UflyMemberStatus == "Elite" | UflyMemberStatus == "Standard", "Me
mber" , "Non-Member"),
         Customers = paste(EncryptedName, birthdateid, sep="_")) %>%
  left_join(Price_Check, by = "PNRLocatorID") %>%
  group_by(Ufly_Flag) %>%
  summarise(num_customers = n_distinct(Customers),
            Num_PNR = n(),
            Revenue = sum(avg_doc_price))%>%
  mutate(Ticket_Per_Customer = round(Num_PNR/num_customers, 2),
         Rev_Per_Customer = round(Revenue/num_customers, 2),
         Rev_Per_PNR = round(Revenue/Num_PNR, 2))
```

```
ggplot(SCA_UFly, aes(x = Ufly_Flag, y = Num_PNR)) +
  geom_bar(stat = 'identity', fill = "#6e0d12") +
  labs(title = "SCA Bookings for Ufly vs non-Ufly Customers",
       y = "# of Bookings", x = "Ufly Membership Status")+
  geom_text(aes(label= scales::comma(Num_PNR)), position = position_stack(vjust = 0.5), color
='gold') +
  theme_bw()
```

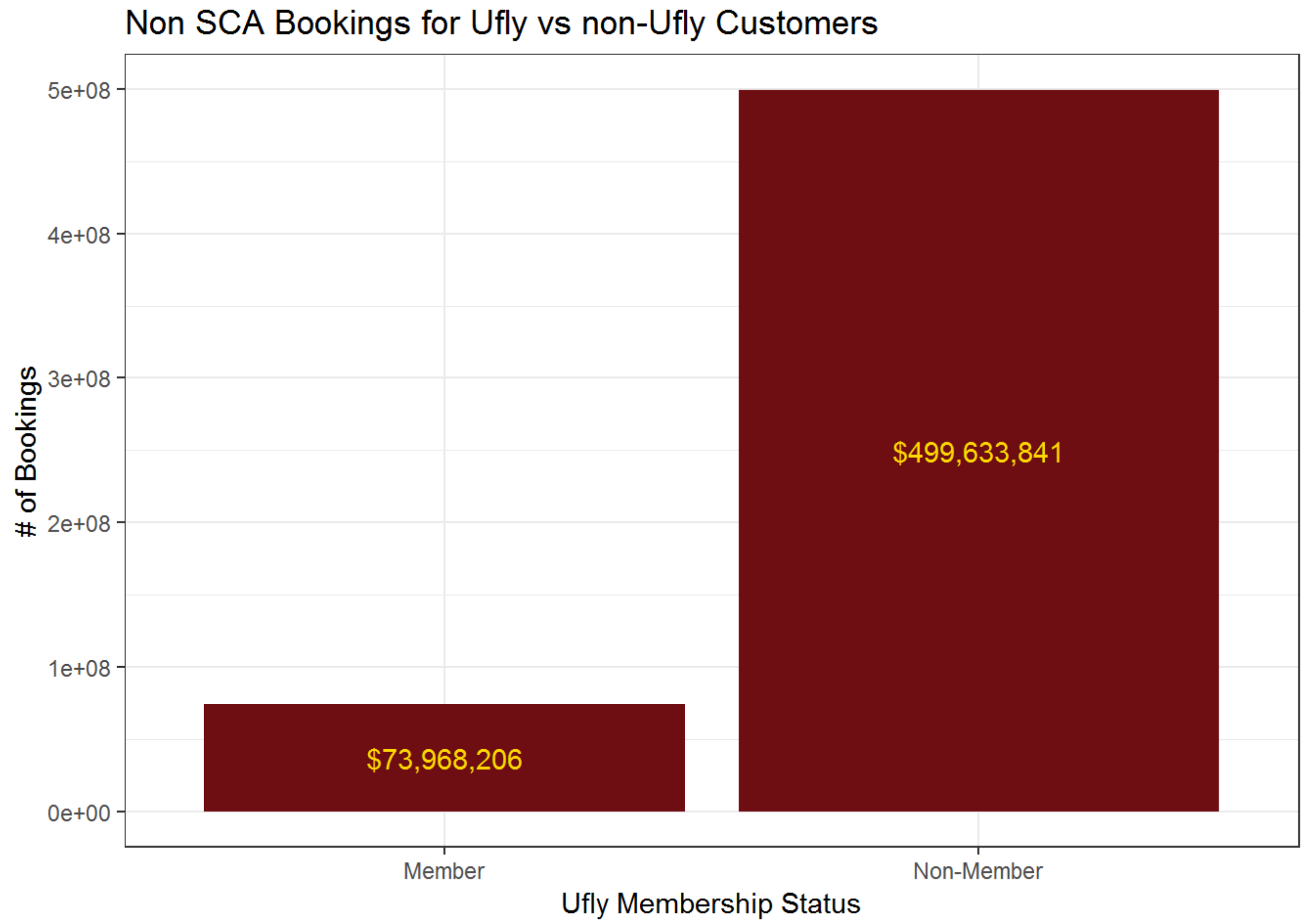


```
ggplot(non_SCA_UFly, aes(x = Ufly_Flag, y = Num_PNR)) +
  geom_bar(stat = 'identity', fill = "#6e0d12") +
  labs(title = "Non SCA Bookings for Ufly vs non-Ufly Customers",
       y = "# of Bookings", x = "Ufly Membership Status")+
  geom_text(aes(label= scales::comma(Num_PNR)), position = position_stack(vjust = 0.5),color=
'gold') +
  theme_bw()
```



From the above graphs, we see that Not all Ufly Members book through the SCA website. The partial revenue generated from these bookings can be considered as a lost opportunity for Sun Country.

```
ggplot(non_SCA_UFly, aes(x = Ufly_Flag, y = Revenue)) +
  geom_bar(stat = 'identity', fill = "#6e0d12") +
  labs(title = "Non SCA Bookings for Ufly vs non-Ufly Customers",
       y = "# of Bookings", x = "Ufly Membership Status")+
  geom_text(aes(label= scales::dollar(Revenue)), position = position_stack(vjust = 0.5), col
 = 'gold') +
  theme_bw()
```



The revenue amounts to be about $73 million over the time period of the data. That is a lot of potential ticket sales which can be driven through the SCA website. Through the general observations we made above to develop our three key points, we will be able to further analyze the data and create thoughtful recommendations to Sun Country. These recommendations will help to direct ticket sales through the SCA website and thus increase revenue.

What is Success?

After our analysis of Sun Countryâ€™s current situation, we define success as quantitatively describing customer segments. Through this, we are helping our client to understand their customer segments and drive booking channel sales within the SCA website. By increasing traffic through the SCA booking website, there will be additional value gained. Better data can be collected on customers which will lead to better customer profiling. The improved customer profiling can produce better target strategy which can lead to an increase in potential Ufly members. All of this additional value can lead to potential revenue increase.

Data Cleaning

Data Understanding and Availability

```
print(colnames(sun_raw))
```

```
##  [1] "PNRLocatorID"         "TicketNum"            "CouponSeqNbr"        
##  [4] "ServiceStartCity"     "ServiceEndCity"       "PNRCreateDate"       
##  [7] "ServiceStartDate"     "PaxName"              "EncryptedName"       
## [10] "GenderCode"           "birthdateid"          "Age"                 
## [13] "PostalCode"           "BkdClassOfService"    "TrvldClassOfService" 
## [16] "BookingChannel"       "BaseFareAmt"          "TotalDocAmt"         
## [19] "UFlyRewardsNumber"    "UflyMemberStatus"     "CardHolder"          
## [22] "BookedProduct"        "EnrollDate"           "MarketingFlightNbr"  
## [25] "MarketingAirlineCode" "StopoverCode"         "Customers"
```

We observe that this is customer booking data, at PNR level. There are different variables like Service Start and End city, Gender, Age, Class of Service Ufly membership, etc. This data can be used to perform consumer analysis to understand different types of customers and their behaviors so that Sun Country can market to them in a better way.

Data Sampling

Considering the size of the data, the team decided to work on a sample of 100,000 customers to generate insights from the data

```
sun_raw$Customers = paste(sun_raw$EncryptedName, sun_raw$birthdateid, sep="_")
sun <- sample(sun_raw$Customers, size = 100000)
Customers <- sun_raw$Customers
```

Upon checking with the original dataset, we found that the data resembled the original sample and the sample was stratified.

Missing Data

Gender

Gender has three options: F, M, and U. Any customer without a gender was assigned U, which we believe stands for unknown. This was done with the code shown below:

```
gender <- sun %>%
  mutate(gender = ifelse(sun$GenderCode == '', 'U',
  ifelse(sun$GenderCode=='F','F', ifelse(sun$GenderCode == 'M', 'M' ,
     ifelse(sun$GenderCode == 'U', 'U', 'U')))))
```

Cardholder

A customer can either be a cardholder or not, but there were a few null entries. We converted all these null entries to False using the code below:

```
CardHolder <- sun %>%
mutate(CardHolder = ifelse(sun$CardHolder == 'true', 'true','false'))
```

Ufly

For the passengers who were neither standard or elite Ufly members, we assigned the value None. The code is shown below:

```
Ufly <- sun %>%
mutate(Ufly = ifelse(sun$UflyMemberStatus == '', 'None',
     ifelse(sun$UflyMemberStatus == 'Elite', 'Elite',
ifelse(sun$UflyMemberStatus == 'Standard', 'Standard', 'NA'))))
```

Stopover

The data include a variable for stopover. For trips which incurred a stopover of less than 24 hours or more than 24 hours we defined accordingly. For trips with a The data include a variable for stopover. For trips which incurred a stopover of less than 24 hours or more than 24 hours we defined accordingly. For trips with a null stopover value we believe to be direct flights. So we decided to identify these trips as none. The code is shown below:

```
layover <- sun %>%
mutate(layover = ifelse(sun$StopoverCode == 'O', 'less_24_hrs', ifelse(sun$StopoverCode ==
'X', 'more_24_hrs','no_stop')))
```

Outliers

Remove Unrealistic Ages

In order to get a dataset representative of the true population, we removed any customers who had unrealistic ages. These were customers who had an age less than zero. We know it is physically impossible for a customer to have an age less than zero, so we removed that observation. Additionally, customers with an age greater than or equal to 100 is unlikely. We know there are very few, if any, customers over the age of 100 who fly. So, we removed these observations as well. This was done with the code below:

```
age <- sun %>% filter(sun$Age > 0 | sun$Age <= 100)
```

Data Transformation and Feature Engineering

Create Unique Customers

In the initial data, there is no unique identifier for each observation. We chose to concatenate the passengerâ€™s encrypted name with their birthdate. We are aware that it is possible for multiple customers to have the same name and date of birth, but this is very unlikely. With a quick check, this was confirmed in our dataset:

```
sun$Customers = paste(sun$EncryptedName, sun$birthdateid, sep="_")
```

Adjusting Gender Observations

Initially in the dataset, there are three different gender identities. Female, Male, and U. We believe U stands for unknown. We also found 43,999 blank values and decided to assign them to the U gender identity. U, because it was not properly defined. This was done with the code below:

```
gender <- sun %>%
  mutate(gender = ifelse(sun$GenderCode == '', 'U',
ifelse(sun$GenderCode=='F','F', ifelse(sun$GenderCode == 'M', 'M' ,
     ifelse(sun$GenderCode == 'U', 'U', 'U')))))
```

Averaged PNR Price

Within the data, each passenger name record (PNR), there are multiple flights and segments rolled into one. This resulted in duplicate prices. First we removed all duplicate prices within each PNR. After this was done, there were still PNR entries with multiple prices. We averaged all the prices together to create one price per PNR entry. This was done with the code below:

```
Price_Check <- sun_raw %>%
  filter(TotalDocAmt != 0) %>%
  group_by(PNRLocatorID, TotalDocAmt) %>%
  summarise(count = n()) %>%
  group_by(PNRLocatorID) %>%
  summarise(avg_doc_price = sum(TotalDocAmt))
```

Booking Channel

The booking channels in the data are divided into ___ unique channels. Because we are only interested in the SCA website versus all other channels. We performed this alteration with the code below:

```
booking_channel <- sun %>%
  mutate(booking_grouped = ifelse(BookingChannel == 'SCA Website Booking','SCA Website Bookin
g','Rest'))
```

Define Variable for Upgrades

Currently, the data provides two separate columns: TrvldClassOfService and BkdClassOfService. By comparing the two columns, we were able to define one column for a customer who receives an upgrade. A value of 0 implies the customer did not receive an upgrade while a value of 1 implies the customer did receive an upgrade. The code is shown below:

```
upgrades <- sun %>%
  mutate(upgraded = ifelse(BkdClassOfService == TrvldClassOfService, 0,1))
```

Remove TotalDocAmt of Zero

In the original data, the TotalDocAmt includes the ticket base fare, taxes, and fees. There are customers who exchange their tickets which is represented by a zero in TotalDocAmt. We can remove any of these entries with the code shown below:

```
Price_Check <- sca %>%
  filter(TotalDocAmt != 0)
```

Cluster Preparation

```
booked_class <- sun %>%
  group_by(Customers, BkdClassOfService) %>%
  summarize(count = n()) %>%
  spread(BkdClassOfService, count, fill =0)
colnames(booked_class) <- c("Customers","b_Coach","b_Discount First Class","b_First Class")


travelled_class <- sun %>%
  group_by(Customers, TrvldClassOfService) %>%
  summarize(count = n()) %>%
  spread(TrvldClassOfService, count, fill =0)
colnames(travelled_class) <- c("Customers","t_Coach","t_Discount First Class","t_First Class"
)

booking_channel <- sun %>%
  mutate(booking_grouped = ifelse(sun$BookingChannel == 'SCA Website Booking','SCA Website Bo
oking','Rest' ))%>%
  group_by(Customers, booking_grouped) %>%
  summarize(count = n()) %>%
  spread(booking_grouped, count, fill =0)

Ufly <- sun %>%
  mutate(Ufly = ifelse(sun$UflyMemberStatus == 'Elite' | sun$UflyMemberStatus == 'Standard',
1, 0))%>%
  group_by(Customers, Ufly) %>%
  summarize(count = n()) %>%
  spread(Ufly, count, fill =0)

Ufly <- Ufly %>% mutate(ufly_flag = ifelse((Elite + Standard)>0, 1, 0))

layover <- sun %>%
  mutate(layover = ifelse(sun$StopoverCode == 'O', 'less_24_hrs',
                          ifelse(sun$StopoverCode == 'X', 'more_24_hrs','no_stop')))%>%
  group_by(Customers, layover) %>%
  summarize(count = n()) %>%
  spread(layover, count, fill =0)

CardHolder <- sun %>%
  mutate(CardHolder = ifelse(sun$CardHolder == 'true', 'true','false'))%>%
  group_by(Customers, CardHolder) %>%
  summarize(count = n()) %>%
  spread(CardHolder, count, fill =0) %>%
  mutate(CardHolder = ifelse(true > 0, 1,0)) %>%
  select(Customers,CardHolder)

age <- sun %>%
  filter(sun$Age > 0 & sun$Age <= 100)%>%
  group_by(Customers) %>%
  summarize(age = max(Age))

gender <- sun %>%
  mutate(gender = ifelse(sun$GenderCode == '', 'U',
                         ifelse(sun$GenderCode == 'F', 'F',
                                ifelse(sun$GenderCode == 'M', 'M' ,
                                       ifelse(sun$GenderCode == 'U', 'U', 'U'))))) %>%
  select(Customers, gender) %>%
  group_by(Customers, gender) %>%
```

```r
  summarize(count = n()) %>%
  spread(gender, count, fill =0)

avg_total_doc_amt <- sun %>%
  group_by(Customers) %>%
  summarize(avg_amt = mean(TotalDocAmt), total_amt = sum(TotalDocAmt), total_tickets = n())

upgrades <- sun %>%
  mutate(upgraded = ifelse(sun$BkdClassOfService == sun$TrvldClassOfService, 0,1))%>%
  group_by(Customers) %>%
  summarize(num_upgrades = sum(upgraded))


early_booked <- sun %>%
  mutate(early_booked = difftime(sun$ServiceStartDate, sun$PNRCreateDate, units = "days"))%>%
  group_by(Customers) %>%
  summarize(avg_difftime = mean(early_booked))

Price_Check <- sun_raw %>%
  filter(TotalDocAmt != 0) %>%
  group_by(PNRLocatorID, TotalDocAmt) %>%
  summarise(num = n()) %>%
  group_by(PNRLocatorID) %>%
  summarise(avg_doc_price = mean(TotalDocAmt))

Price_customers <- sun_raw %>%
  filter(TotalDocAmt != 0) %>%
  group_by(PNRLocatorID, Customers) %>%
  summarise(num_customers = n()) %>%
  group_by(PNRLocatorID) %>%
  summarise(num_customers = n())

Customers_PNR <- Price_Check %>%
  left_join(Price_customers, by = 'PNRLocatorID') %>%
  mutate(avg_ticket_price = avg_doc_price/num_customers)%>%
  left_join(sun_raw, by = 'PNRLocatorID') %>%
  select(PNRLocatorID, avg_doc_price, num_customers, avg_ticket_price, Customers) %>%
  group_by(Customers) %>%
  summarise(ticket_price = sum(avg_ticket_price))

customers <- sun %>% distinct(Customers)
SC_final <- merge(customers, age, by = "Customers", )
SC_final <- left_join(customers, age, by = c("Customers") )
SC_final <- left_join(SC_final, avg_total_doc_amt, by = c("Customers") )
SC_final <- left_join(SC_final, booked_class, by = c("Customers") )
SC_final <- left_join(SC_final, booking_channel, by = c("Customers") )
SC_final <- left_join(SC_final, CardHolder, by = c("Customers") )
SC_final <- left_join(SC_final, gender, by = c("Customers") )
SC_final <- left_join(SC_final, layover, by = c("Customers") )
SC_final <- left_join(SC_final, travelled_class, by = c("Customers") )
SC_final <- left_join(SC_final, Ufly, by = c("Customers") )
SC_final <- left_join(SC_final, upgrades, by = c("Customers") )
SC_final <- left_join(SC_final, early_booked, by = c("Customers") )
SC_final <- left_join(SC_final, Customers_PNR, by = c("Customers"))
```

Data Normalization:

Since we will perform distance based clustering, we normalize the variables using standard scaling. We do this so that outliers do not affect the variables which are more centrally distributed in the data. And we also keep outliers (customers which have high spend or book in high amount) on the website in the given data timeframe.

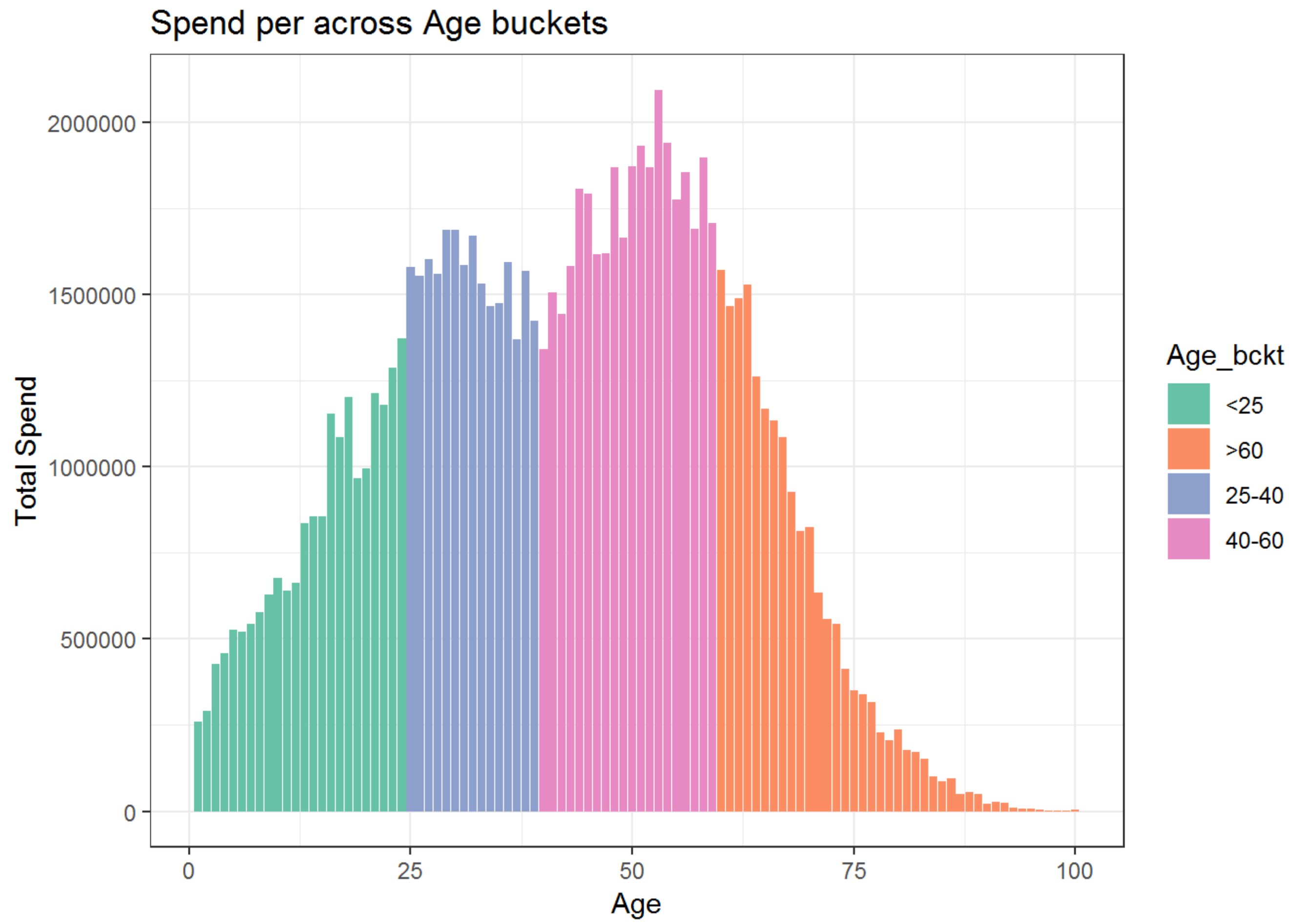
```
normalize <- function(x){
  return ((x - mean(x, na.rm = T))/sd(x, na.rm = T))}
```

Age:

For this variable, we observe the distribution of customer spends across different ages. The motive is to find groups of customers with similar spending patterns which can then be classified into similar buckets.

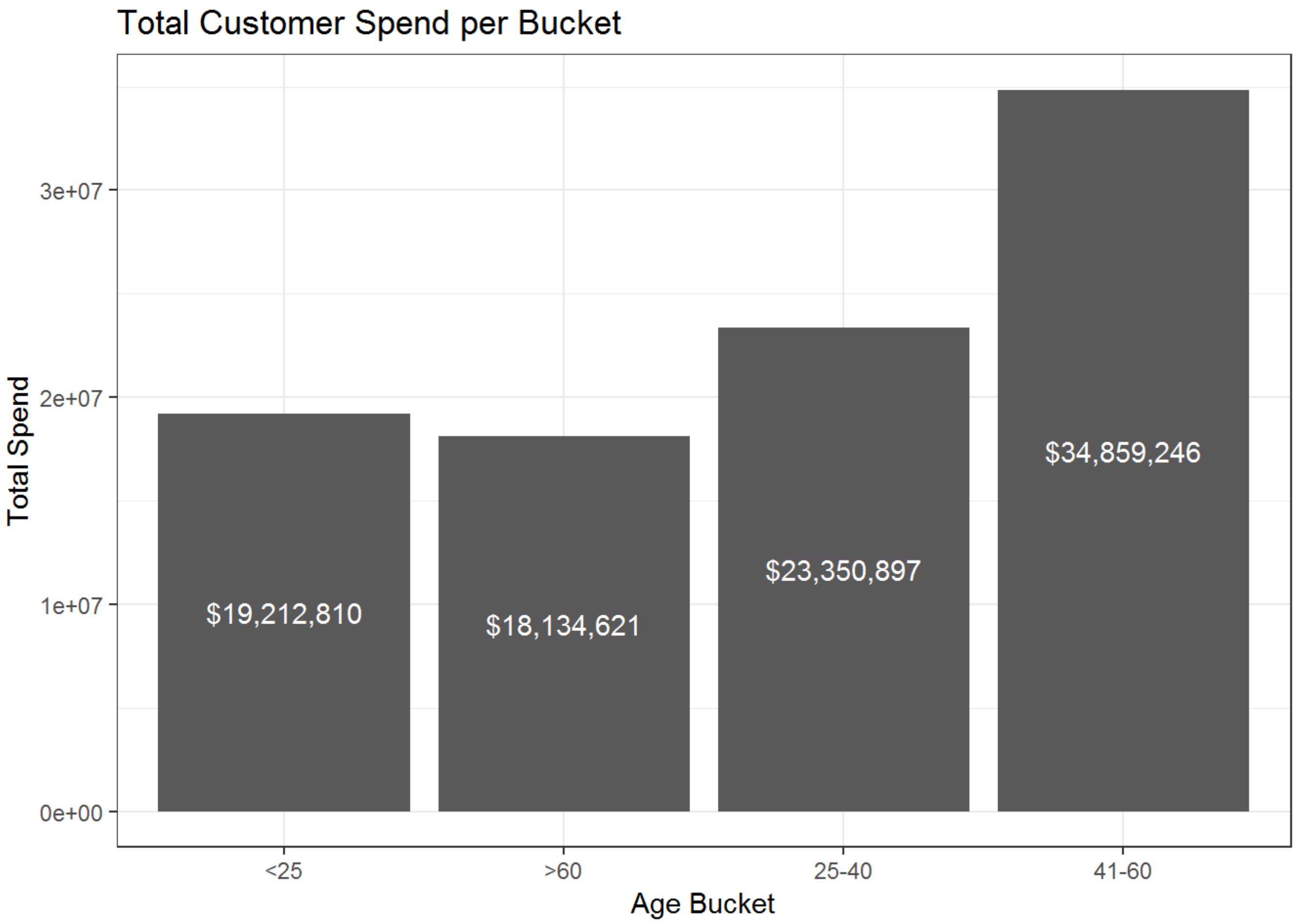
```
age_income <- SC_final %>%
  group_by(age) %>%
  summarise(avg_amt = sum(total_amt))%>%
  mutate(Age_bckt = ifelse(age<25, "<25", ifelse(age<40, "25-40", ifelse(age<60, "40-60", ">6
0"))))

age_income2 <- age_income %>%
  na.omit() %>%
  mutate(age_bucket = ifelse(age<25, "<25", ifelse(age<40, "25-40", ifelse(age<60, "41-60",
">60")))) %>%
  group_by(age_bucket) %>%
  summarise(avg_amt = sum(avg_amt))

age_income3 <- age_income %>%
  na.omit() %>%
  mutate(age_bucket = ifelse(age<25, "<25", ifelse(age<40, "26 - 40", ifelse(age<60, "40-60",
">60"))))
```

```
ggplot(age_income, aes(x = age, y = avg_amt, fill = Age_bckt)) +
  geom_bar(stat = 'identity')+
  labs(title='Spend per across Age buckets', x = 'Age', y = 'Total Spend')+
  scale_fill_brewer(palette="Set2") +
  theme_bw()
```



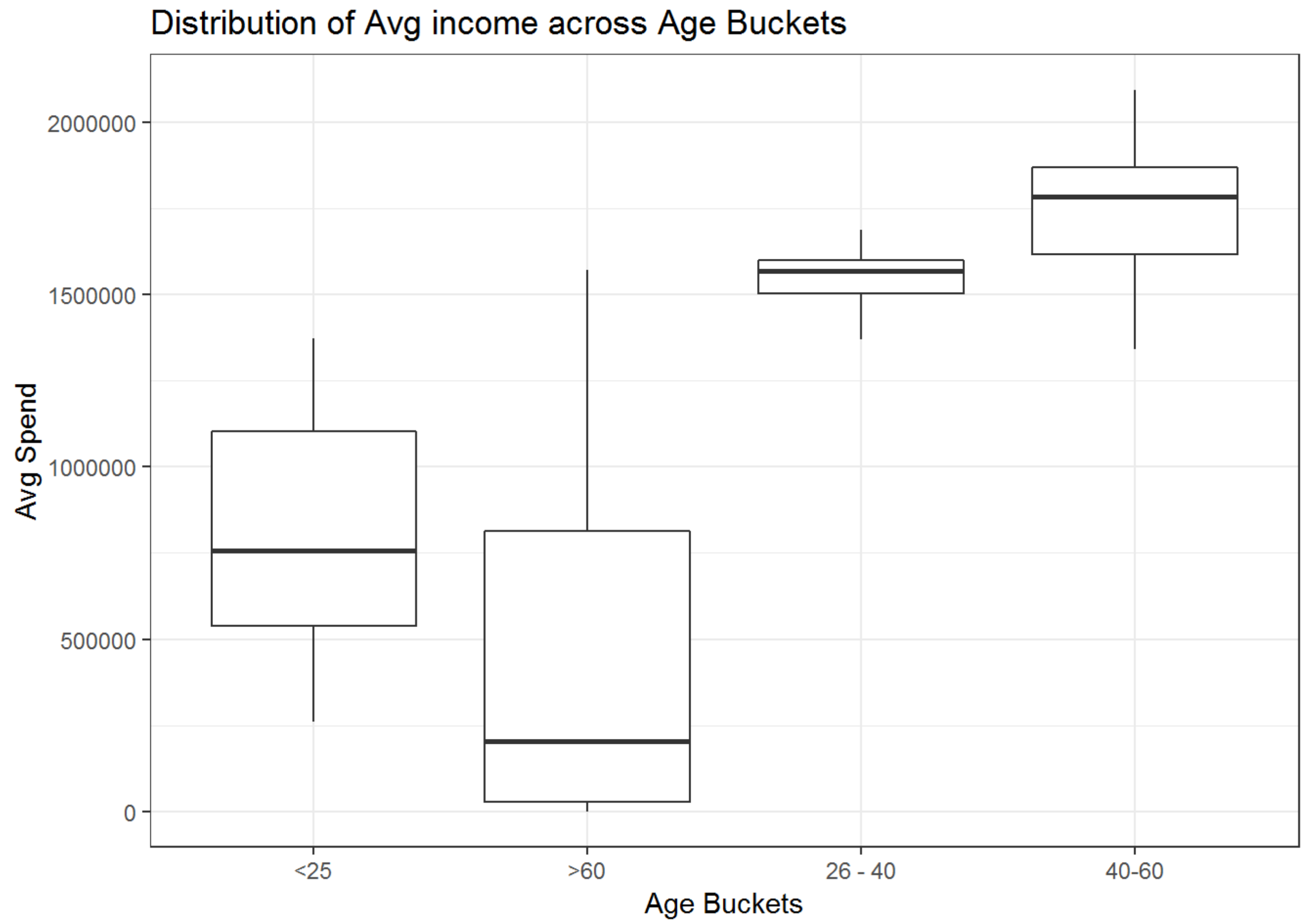
From the above plot, we observe that a customer can fall into one of four distinct buckets: - Low Spenders: 1) Consumers who are less than 25 years of age, excluding 25 2) Customers who are greater than 60 years of age, excluding 60 - High Spenders: 1) Customers between the ages of 25 and 40 2) Customers between the ages of 41 and 60 We can also confirm their age groups' total spending below:

```
ggplot(age_income2, aes(x = age_bucket, y = avg_amt)) + geom_bar(stat = 'identity')  +
  labs(title = "Total Customer Spend per Bucket", x = "Age Bucket", y = "Total Spend") +
  geom_text(aes(label= scales::dollar(avg_amt)),position = position_stack(vjust = 0.5), color
= 'white') +
  theme_bw()
```



Next, we look at the boxplots for distribution which show the spread of income within the age groups.

```
ggplot(age_income3, aes(x = age_bucket, y = avg_amt)) +
  geom_boxplot() +
  labs(title = "Distribution of Avg income across Age Buckets",
       x="Age Buckets", y="Avg Spend") +
  theme_bw()
```



We encode the buckets as following: 0: Customers older than 60 years and younger than 25 years as they have similar spending patterns. 1: Customer in 25-40 years 2: Customer between 41-60 years of age

Analysis

Description and Rationale for the Chosen Analysis

The best way to increase booking sales on the SCA website is to determine which types of customers prefer to book on the SCA website over the non-SCA website. To understand the customers by finding groups of similar customers, we use clustering, which allows us to not only find similar groups, but also allows to compare differences between various customer segments which can drive customer targeting strategy to drive revenue. We clustered Sun Country customers based on the following variables:

For basic customer understanding, we select the following variables for clustering.

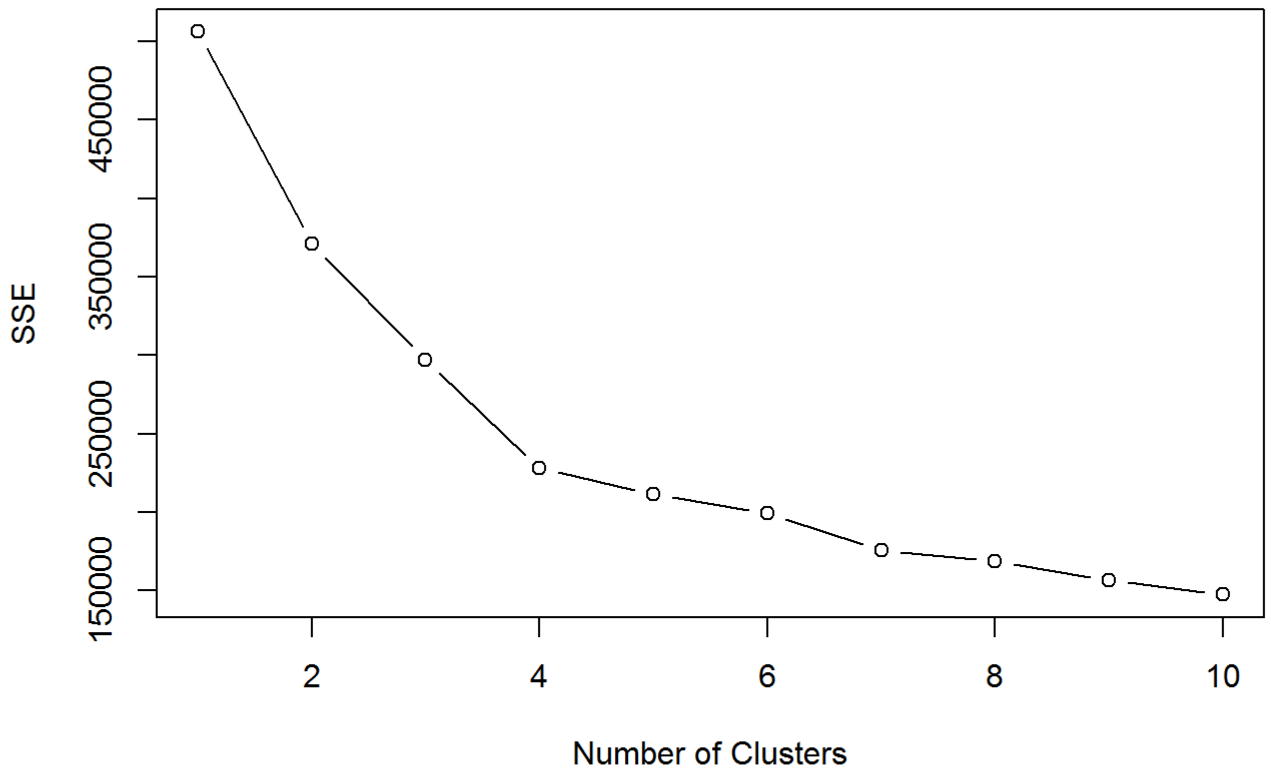
1. Age of customer
2. Total amount of spent on tickets per customer (normalized)
3. Total number of tickets per customer
4. Means of booking ticket per customer
5. Sun Country card status
6. Number of upgrades per customer
7. Ufly status per customer

Clustering:

We run K-Means algorithm on the cleaned data to generate clusters. We first plot the SSE curve to find optimal number of clusters:

```
colnames(SC_final)
cols <- c('Customers','age', 'avg_amt', 'total_amt', 'total_tickets', 'SCA Website Booking',
'CardHolder', 'num_upgrades', 'avg_difftime', 'more_24_hrs', "ufly_flag", 'avg_ticket_price')
sun_data <- SC_final[cols]

sun_scale <- sun_data %>%
  mutate(age_bucket = ifelse(age<25, 0, ifelse(age<40, 1, ifelse(age<60, 2, 0))),
         avg_amt = normalize(avg_amt),
         norm_total_amt = normalize(avg_amt),
         total_tickets = normalize(total_tickets),
         bookings = normalize(`SCA Website Booking`),
         upgrades = normalize(num_upgrades),
         difftime = normalize(as.numeric(avg_difftime)),
         more_24 = normalize(more_24_hrs),
         Card = CardHolder) %>%
  select(Customers, age_bucket,norm_total_amt, total_tickets,bookings, Card, upgrades, ufly_f
lag) %>%
  na.omit()

set.seed(123)
SSE_curve <- c()
for (k in 1:10) {
  kcluster <- kmeans(sun_scale[-c(1)], k)
  sse <- sum(kcluster$withinss)
  SSE_curve[k] <- sse
}
```

```
plot(1:10, SSE_curve, type="b", xlab="Number of Clusters", ylab="SSE")
```



Looking at the SSE curve, we find the optimal number of clusters to be 4. We proceed to make 4 clusters from the data, and then analyze them to find characteristics.

```
kcluster <- kmeans(sun_scale[-c(1)], 4)

# Profiles
sun_scale <- rbind(sun_scale, kcluster$cluster)
clust_cust <- c(sun_scale$Customers)
SC <- SC_final %>% filter(Customers %in% clust_cust)
SC <- merge(SC, sun_scale, by = "Customers", all.x=TRUE)


Clust <- SC %>%
  rename(Cluster = "kcluster$cluster", SCA_Booking = "SCA Website Booking", total_price = "ti
cket_price" ) %>%
  mutate(elite_flag = ifelse(Elite>0,1,0),
         standard_flag = ifelse(Standard>0,1,0),
         male_flag = ifelse(M>0, 1, 0),
         female_flag = ifelse(`F`>0, 1, 0),
         price_per_ticket = total_price/total_tickets.x,
         SCA_flag = ifelse(SCA_Booking>0, 1, 0)
  ) %>%
  group_by(Cluster) %>%
  summarise(Cust_num = n(),
            med_age = median(age),
            how_early = mean(avg_difftime),
            avg_upgrades = mean(num_upgrades),
            perc_CardHolder = mean(CardHolder),
            avg_ticket_price = mean(price_per_ticket, na.rm = T),
            Total_amt = sum(total_price, na.rm = T),
            SCA_Booking = sum(SCA_Booking),
            SCA_Booking_cnt = sum(SCA_flag),
            Other_Booking = sum(Rest),
            Female = sum(`F`),
            Male = sum(M),
            Female = sum(male_flag),
            Male = sum(female_flag),
            elite = sum(Elite),
            standard = sum(Standard),
            elite_cnt = sum(elite_flag),
            standard_cnt = sum(standard_flag),
            Ufly = sum(ufly_flag.x)
  ) %>%
  mutate(other_perc = Other_Booking / (SCA_Booking + Other_Booking),
         Cluster_Names = ifelse(Cluster == 2, "Potential Devotees",
                                ifelse(Cluster == 3, "Aspiring Loyals",
                                       ifelse(Cluster == 4, "Forever Unwavering",
                                              ifelse(Cluster == 1, "Trusty Passengers", ""
)))))
```

From this, we were able to identify four different types of customers:

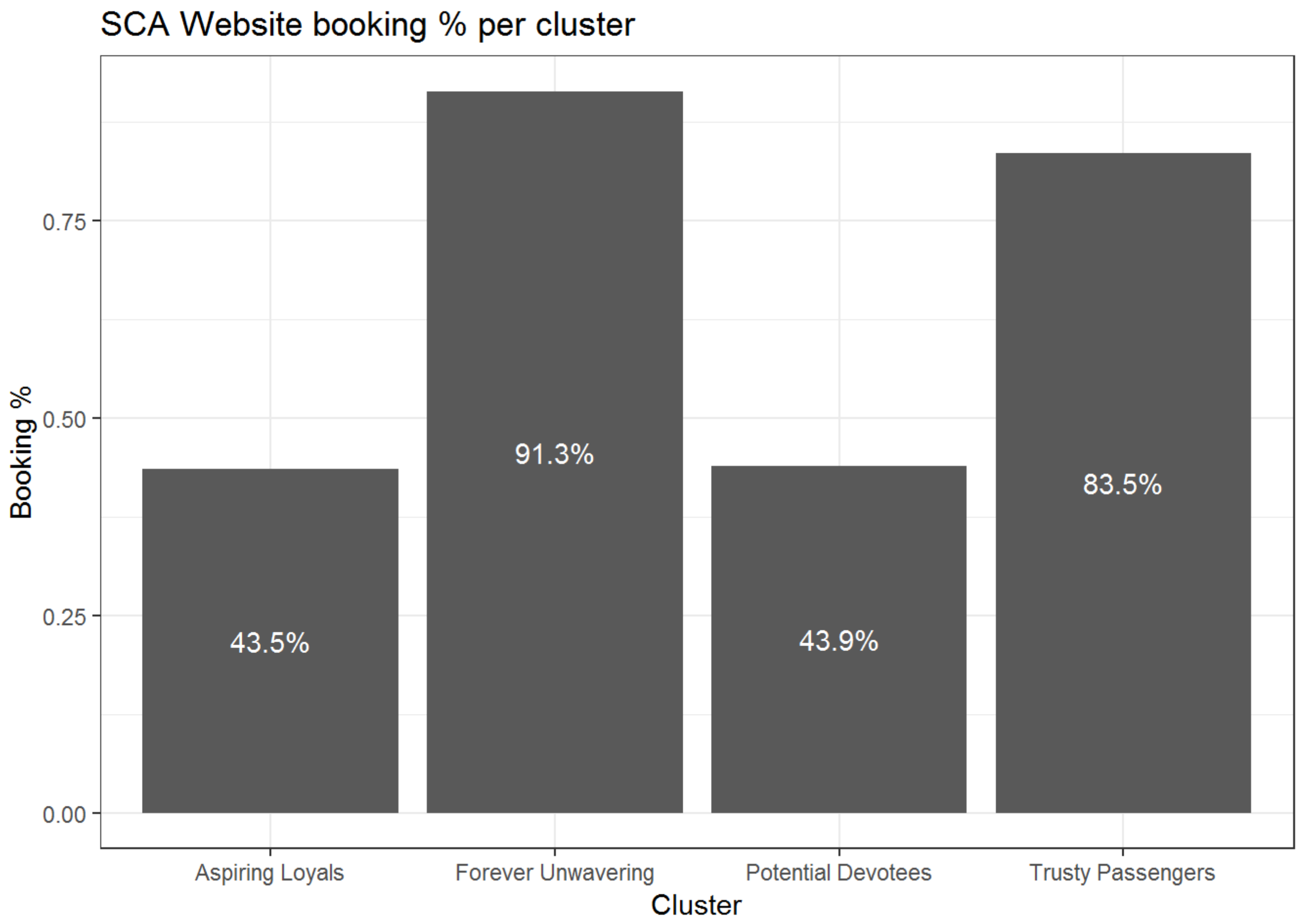
1. Forever Unwavering (high likelihood of booking through SCA website) (0.2% : 219 Customers)
2. Trusty Passengers (5.8% : 5544 Customers)
3. Potential Devotees (31% : 29,543 Customers)
4. Aspiring Loyals (low likelihood of booking through SCA) (63% : 60,055 Customers)

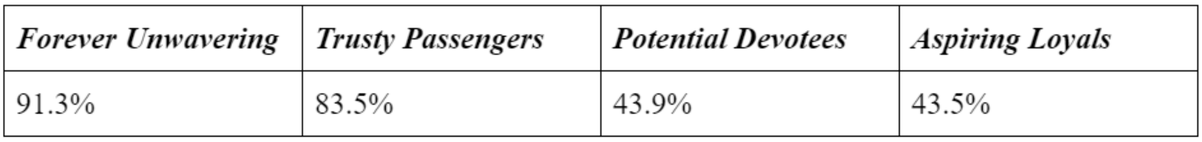
From these four groups, we will be able to effectively draw conclusions within the groups and how to best increase traffic on the SCA booking site.

Group Characteristics

The first thing we were interested in observing would be what percentage of customers in each group purchase their tickets through the SCA website and how much they spend. We obtain the following information using the code listed above:

```
ggplot(Clust, aes(x = Cluster_Names, y = SCA_Booking_cnt / Cust_num)) +
  geom_bar(stat="identity") +
  labs(title = "SCA Website booking % per cluster", x = "Cluster", y = "Booking %") +
  geom_text(aes(label= scales::percent(SCA_Booking_cnt / Cust_num)),
            position = position_stack(vjust = 0.5), color = 'white') +
  theme_bw()
```



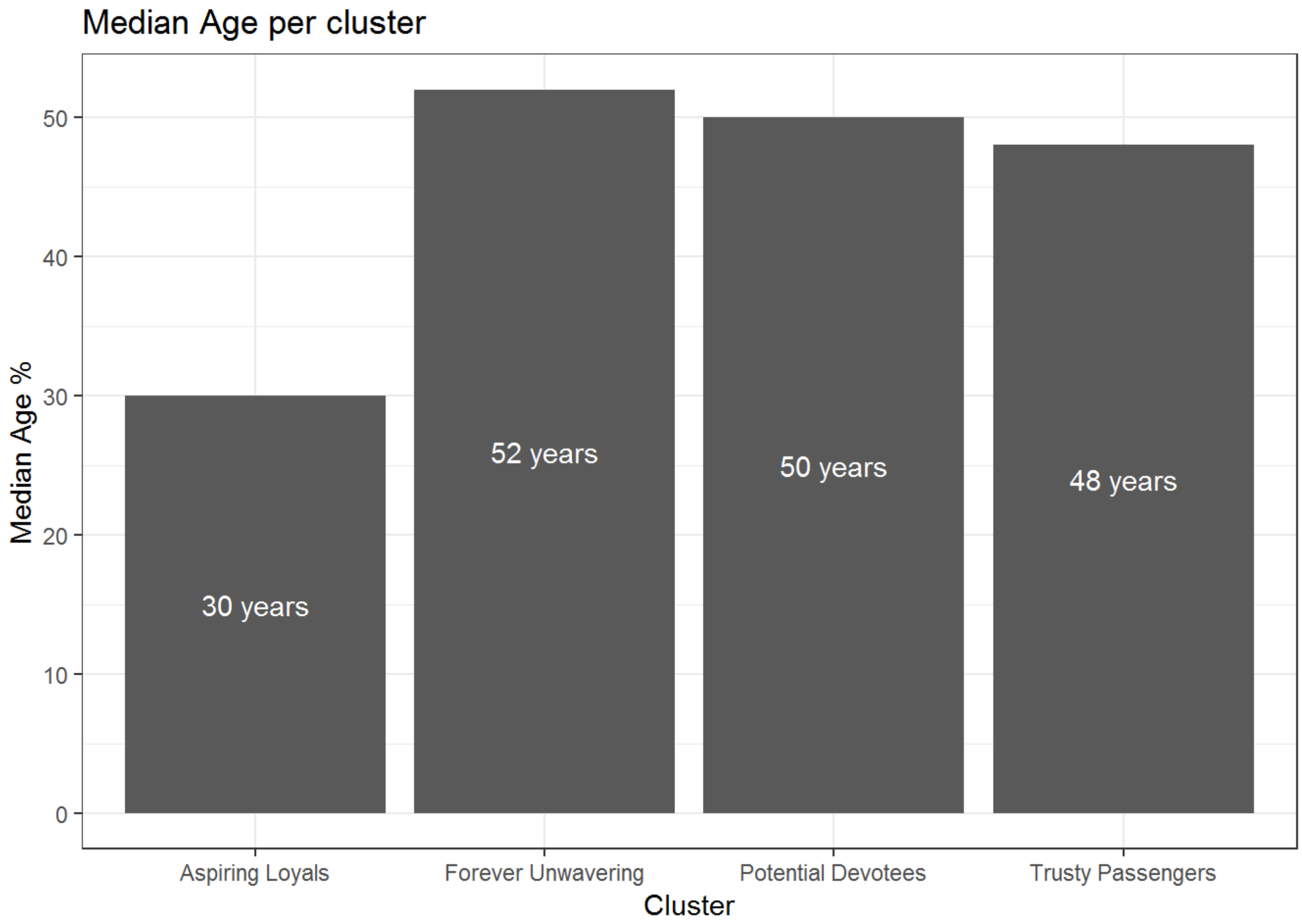


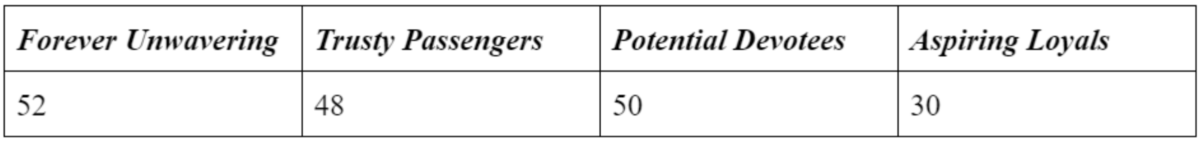
Forever Unwavering	*Trusty Passengers*	*Potential Devotees*	*Aspiring Loyals*
91.3%	83.5%	43.9%	43.5%

Cluster SCA Booking Percentage

Within these four groups, the Forever Unwavering and Trusty Passengers are consistent in their booking through the SCA website. The remaining groups, Potential Devotees and Aspiring Loyals, tend to purchase fewer tickets through the SCA website. Upon further inspection, we observed the median age for each group. We obtain the following information using the code listed above:

```
ggplot(Clust, aes(x = Cluster_Names, y = med_age)) +
  geom_bar(stat="identity") +
  labs(title = "Median Age per cluster", x = "Cluster", y = "Median Age %") +
  geom_text(aes(label= paste(med_age, "years")), position = position_stack(vjust = 0.5), colo
r = 'white') +
  theme_bw()
```





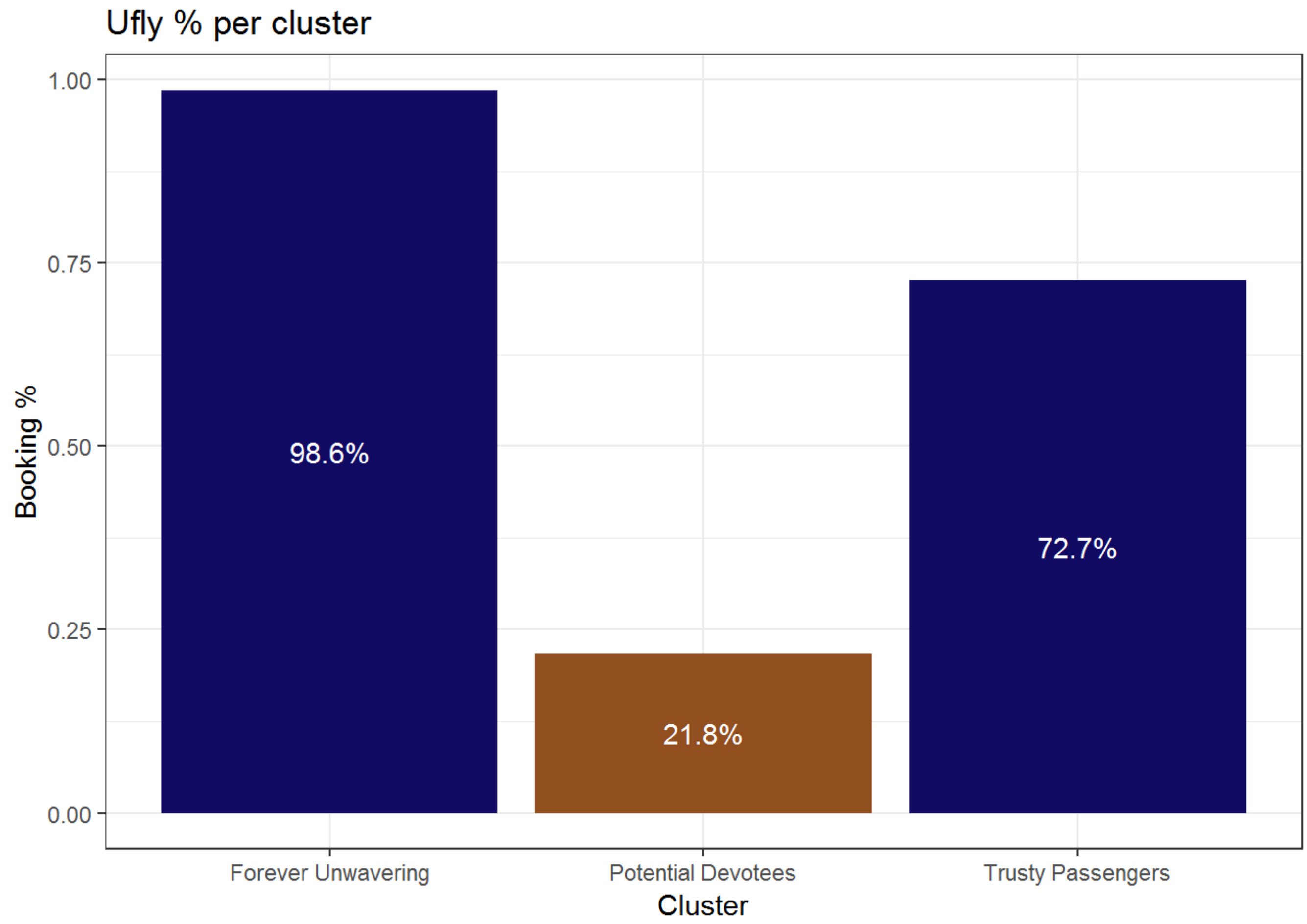
Forever Unwavering	***Trusty Passengers***	***Potential Devotees***	***Aspiring Loyals***
52	48	50	30

Cluster Median Age

Based on this information, we can see that the Forever Unwavering, Trusty Passengers, and Potential Devotees are all similar in age. We can also see that the Aspiring Loyals have a median age of 20 years younger than the previous group.

Finally, we would like to see what distinguishes the Forever Unwavering, Trusty Passengers, and Potential Devotees from one-another. We noticed the percentage of Ufly members serves as an important distinction:

```
ggplot(Clust[c(1,2,4),], aes(x = Cluster_Names, y = Ufly / Cust_num)) +
  geom_bar(stat="identity", fill = c('#110a63', '#914e1f', '#110a63')) +
  labs(title = "Ufly % per cluster", x = "Cluster", y = "Booking %") +
  geom_text(aes(label= scales::percent(Ufly / Cust_num)),
            position = position_stack(vjust = 0.5), color = 'white') +
  theme_bw()
```



Forever Unwavering	*Trusty Passengers*	*Potential Devotees*
98.6%	72.7%	21.8%

Cluster UFly Membership

While Potential Devotees appears on the surface to be very similar to Forever Unwavering and Trusty Passengers, we can now see that is not the case. While all three groups have the same age, there is a distinct division in percentage of Ufly members. While Forever Unwavering and Trusty Passengers both have a large percentage of Ufly members, the Potential Devotees have a significantly smaller percentage.

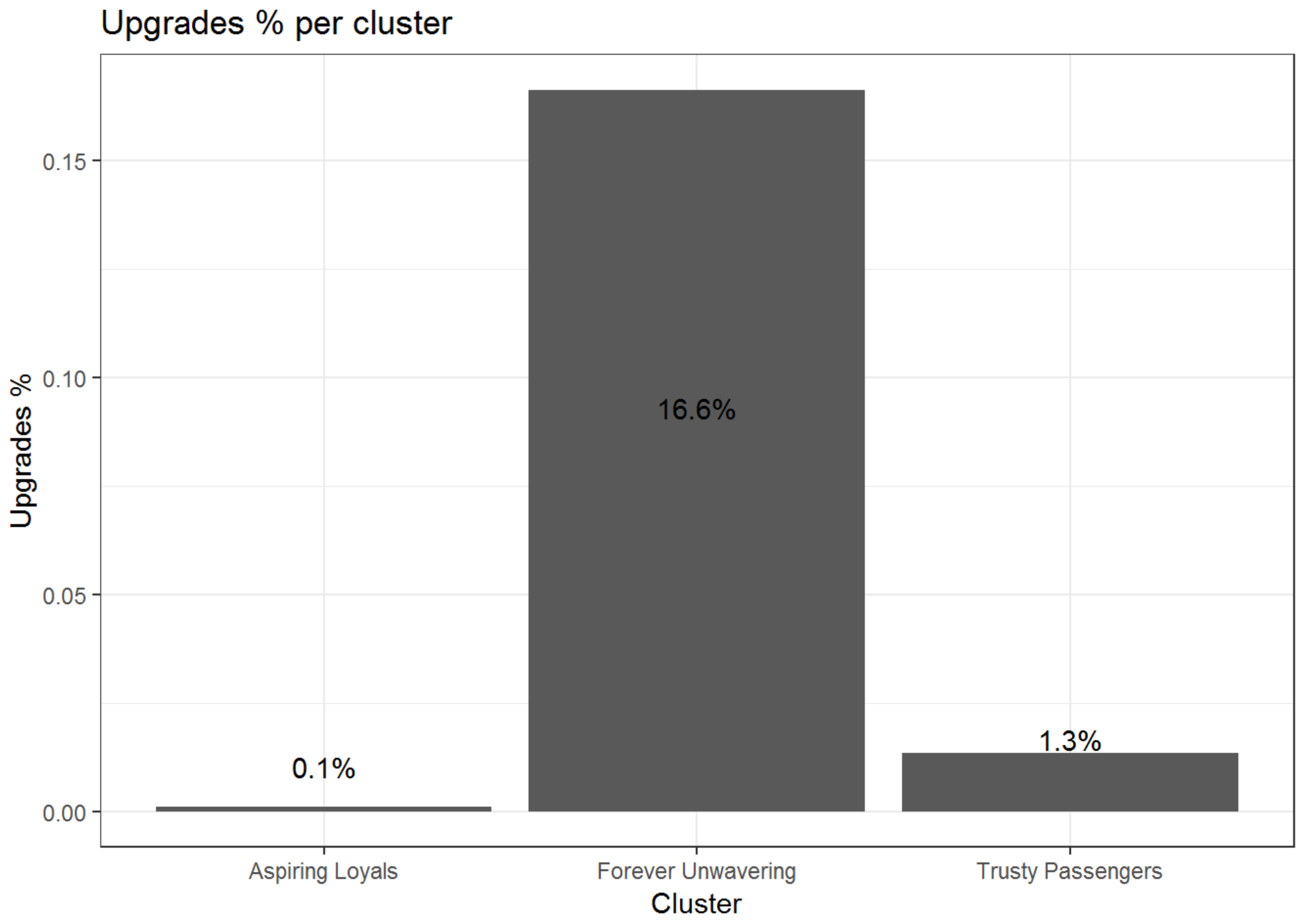
It is through these general observations, we can perform further investigation to derive solutions.

Investigation into Potential Devotees \

As we have previously seen, Potential Devotees is very similar to Forever Unwavering and Trusty Passengers. They already have the same age, but the groups differ in Ufly status. If we could increase the Ufly status percentage of the Potential Devotees, they would become more similar to either the Forever Unwavering or the Trusty Passengers. In turn, SCA booking may increase as well for the Potential Devotees.

In order to find the best method to increase the Ufly percentage of the Potential Devotees, we analyzed information that may be related to Ufly status. The likelihood of upgrades appeared to be the most useful. We obtain the following information using the code listed above:

```
ggplot(Clust[c(1,3,4),], aes(x = Cluster_Names, y = avg_upgrades/100)) +
  geom_bar(stat="identity") +
  labs(title = "Upgrades % per cluster", x = "Cluster", y = "Upgrades %") +
  geom_text(aes(label= scales::percent(avg_upgrades/100)),
            position = position_stack(vjust = 0.5), color = 'black', vjust = -1.5) +
  theme_bw()
```



Forever Unwavering	*Potential Devotees*	*Trusty Passengers*
16.6%	0.15%	1.34%

Cluster : Average Upgrades

This information shows that with a Ufly membership comes an increase likelihood of an upgrade. An upgrade would be a change in seating class from the initial booking of tickets and the actual flight. For example, a passenger being upgraded from coach to first class. We can use this insight to create a solution that would be beneficial to Sun Country.

Investigation into Aspiring Loyals

Returning to the Aspiring Loyals, recall the age for this group is about twenty years younger than the remaining groups. This requires a different analysis than the Potential Devotees. Because of the vast age difference between the groups, there is likely an effect this has on booking patterns. We know the Aspiring Loyals are of the age to be millenials. When millennials make the purchase decision, the winning factor is price[1]. This is reflected in the higher use of non-SCA booking websites. Using this information, we can develop a solution that would shift this attention from the non-SCA booking websites to the official SCA booking channel.

Final Takeaways

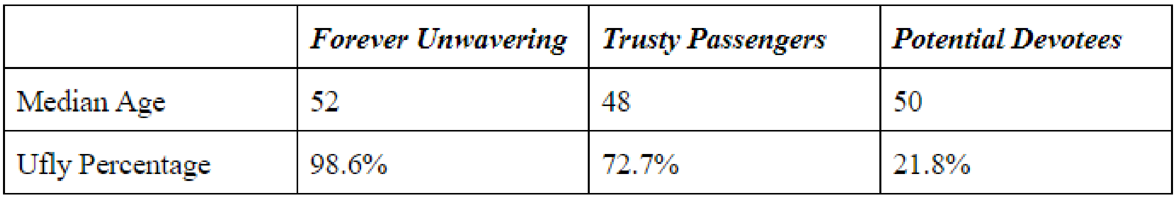
By dividing the consumers into 4 major segments, customer profiles for Sun Country can be defined to further develop marketing strategies. Through these segments, we can more effectively convert customers from non-SCA booking websites to the SCA booking website. From our clustering, we developed two strategies were developed based on observing the consumer profiles:

1. Millennials can be reached via new avenues of engagement, namely social media [2]. Thus in order to increase SCA website traffic, Sun Country could advertise on social media.

2. Increase the amount of Ufly memberships within the Potential Devotees through advertising the increased likelihood of upgrades.

The first action is fairly self explanatory. Advertising on social media will appeal to the members of the Aspiring Loyals group and encourage them to overlook the cheap prices available on non-SCA booking sites. This will increase SCA website traffic.

The second action is a little more cryptic. It is very difficult to find a passenger that will be uninterested in upgrades. Properly informing the Potential Devotees of the increased chance of an upgrade by becoming a Ufly member may encourage them to sign up for a membership. When the percentage of Ufly membership increases, the Potential Devotees group will look more similar to the Trusty Passengers and Forever Unwavering. This is because if you look at the table below, Potential Devotees already has the same age as Trusty Passengers and Forever Unwavering. Increasing the Ufly percentage will make the Potential Devotees look very similar to existing groups.



	Forever Unwavering	*Trusty Passengers*	*Potential Devotees*
Median Age	52	48	50
Ufly Percentage	98.6%	72.7%	21.8%

So it will become likely for the customers in the Potential Devotees group to merge with either the Trusty Passengers or the Forever Unwavering. These consumers would likely adopt the same characteristics of their new groups. This includes an increased booking through the SCA website.

These two actions can effectively increased the traffic on SCA website. Additionally, in the long run these actions may bring many benefits such as better data collected, enhanced consumer profiling, personalized target strategy, and increased potential Ufly members. All of these future benefits in the long run can increase Sun Countrys revenue.

Reference

1. Crawford, K. (n.d.). The millennial consumer: how they shop & why they buy: Blog: Ascend. Retrieved from https://www.herosmyth.com/article/millennial-consumer-how-they-shop-why-they-buy.\newline (https://www.herosmyth.com/article/millennial-consumer-how-they-shop-why-they-buy.\newline)

2. 6 Ways Millennials are Redefining Customer Service. (n.d.). Retrieved from https://www.salesforce.com/blog/2017/08/how-millennials-are-redefining-customer-service.html.\newline (https://www.salesforce.com/blog/2017/08/how-millennials-are-redefining-customer-service.html.\newline)

3. Carnoy, J. (2017, March 3). 5 Ways Social Media Has Transformed Tourism Marketing. Retrieved from https://www.entrepreneur.com/article/286408 (https://www.entrepreneur.com/article/286408).