Homework 2: Sun Country - Marketing in the Digital Age of the Airline Industry

Adithya Ramesh, Jie Liu, Brady Engelke, Priya Priya, Sharang Jindal, Zixuan Wang

10/20/2019

Contents

0.1 Situation Overview

Sun Country Airlines (SCA) has established itself as a formidable airline, particularly in the Minneapolis market. Executives at SCA are proud of this accomplishment, yet aware that innovative management strategies need to be developed if they hope to continue the company's prosperity in the highly competitive airline industry. SCA does not have the luxury to compete on fare-price as larger airlines do. Thus, creative ways of delivering value to customers are imperative to the success of SCA's business model. Michael Warnken, Senior Director of e-Commerce, and Roselie Vaughn, Director of Customer Digital Experience, both play pivotal roles in driving business initiatives to accomplish this task. They have contacted the Carlson Consulting Team to extract insights from the data SCA collects on its flights/customers in hopes the consultants can define SCA's customer behavior in a way that they can use to develop data-driven marketing strategies.

The Carlson Consulting Team has broken down this task into three objectives:

1. Provide a clear picture of the behavior of SCA's customers
2. Identify unique groups of customers within SCA's overall customer set
3. Leverage similar customer groups and the behavior of customers to derive fresh marketing strategies

After the reconciliation of findings, the Carlson Consulting Team proposes two opportunities for Warnken and Vaughn to develop marketing strategies around. If exploited, these recommendations will provide a more intimate vacation experience for customers, improve the effectiveness of SCA's promotional products, and boost SCA's bottom line over time.

This document illustrates how the Carlson Consulting Team arrived at these recommendations and elaborates on the value these recommendations can bring to SCA.

0.2 Where, how, and when do customers fly with SCA?

Execution and Results: First, the raw data needed to be cleaned and transformed into different levels of granularity. This step derived the foundational dataframes our analyses are built upon.

Below is the level of granularity for each dataframe: 1. Each row corresponding to a single flight 2. Each row corresponding to an entire trip 3. Each row corresponding to a unique customer

Load necessary packages and clean raw dataset

```
library(tidyverse)
library(cluster)
library(factoextra)
library(lubridate)
library(naniar)
library(ggpubr)
library(Rtsne) # for t-SNE plot
library(clustMixType)
```

Remove rows with faulty Gendercode and BirthdateID

```
data <- data %>%filter(!is.na(birthdateid))
data$GenderCode<-as.character(data$GenderCode)
data <- data %>%filter(GenderCode!="")
data$GenderCode<-as.factor(data$GenderCode)
```

Replace faulty values in Age with median value

```
data$Age[data$Age < 0] <- median(data$Age)
data$Age[data$Age > 100] <- median(data$Age)
```

Replace Missing values in UflyMemberStatus with "non-ufly"

```
data$UFlyRewardsNumber[is.na(data$UFlyRewardsNumber)] <- 0
data$UflyMemberStatus<-as.character(data$UflyMemberStatus)
data$UflyMemberStatus[data$UflyMemberStatus==''] <-"non-ufly"
```

Replace rows with faulty city codes as BookingChannel with "Other"

```
data$BookingChannel<-as.character(data$BookingChannel)
data$BookingChannel[data$BookingChannel!="Outside Booking" &
                      data$BookingChannel!="SCA Website Booking" &
                      data$BookingChannel!="Tour Operator Portal" &
                      data$BookingChannel!="Reservations Booking" &
                      data$BookingChannel!="SY Vacation"] <- "Other"
data$BookingChannel<-as.factor(data$BookingChannel)
```

Remove rows with MarketingAirlineCode codes other than "SY". Data for other airlines is not applicable to Sun Country's customer profile

```
data$MarketingAirlineCode <- as.character(data$MarketingAirlineCode)
data <- data %>% filter(MarketingAirlineCode=="SY")
data$MarketingAirlineCode <- as.factor(data$MarketingAirlineCode)
```

Creating a new column called error that refers to customers who do not start with coupon sequence number 1. We want to exclude customers who are gaming the system of purchasing tickets and not fulfilling all of the trips on the PNR

```
data <- data %>% group_by(PNRLocatorID) %>%
  mutate(error = ifelse(min(CouponSeqNbr)!= 1, 1, 0)) %>%
```

```r
  filter(error == 0) %>%
  ungroup()
```

create uid

```r
# create uid based on encryptedname, gender and birthdate
df_uid <- data4 %>% group_by(EncryptedName, GenderCode, birthdateid) %>% summarise(n())
df_uid$uid <-  seq.int(nrow(df_uid))
data4 <- merge(data4, df_uid, by=c("EncryptedName", 'GenderCode', 'birthdateid'))

# now Let's drop EncryptedName and n(), error
sc_id <-  subset(data4, select = -c(`EncryptedName`, `n()`, `error`))

# save the output
# write.csv(sc_id, file = 'sc_withid.csv', row.names = FALSE)
```

The above code snips were performed on a different script in order to create sc_withuid.csv which is read in below. This had to be done due to the sheer quantity of raw data and time constraints

```r
# read cleaned suncountry original data with uid
sc_id <- read.csv('sc_withuid.csv', header = TRUE)
```

```r
uniqueUids <- unique(sc_id$uid)
# To produce the same sample every time the code is run
# get 10000 unique travelers
set.seed(111)
sample_Uids <- sample(uniqueUids, 10000)
sample_PNRs <- unique(sc_id %>%
                        filter(uid %in% sample_Uids) %>%
                        select(PNRLocatorID))

data <- sc_id %>%
  filter(PNRLocatorID %in% sample_PNRs$PNRLocatorID)
```

```r
data <- data %>%
  group_by(PNRLocatorID) %>%
  mutate(error = ifelse(min(CouponSeqNbr)!= 1, 1, 0)) %>%
  filter(error == 0) %>%
  ungroup()
```

Feature engineering. First, create season and month cols

```r
data$ServiceStartDate <- as.Date(data$ServiceStartDate)
data <- data %>%
  mutate(month = month(ServiceStartDate)) %>%
  mutate(Season = ifelse(month %in% c(1, 2, 12), 'winter',
                  ifelse(month %in% c(3, 4, 5), 'spring',
                  ifelse(month %in% c(6, 7, 8), 'summer', 'fall'))))

data$PNRCreateDate <- as.character(data$PNRCreateDate)
data$ServiceStartDate <- as.character(data$ServiceStartDate)
```

Transform book & travel class of service into numerical. Create upgraded col

```r
data <- data %>%
  mutate(book_class = ifelse(BkdClassOfService == 'First Class', 3,
                      ifelse(BkdClassOfService == 'Discount First Class', 2, 1))) %>%
```

```r
  mutate(travel_class = ifelse(TrvldClassOfService == 'First Class', 3,
                        ifelse(TrvldClassOfService == 'Discount First Class', 2, 1))) %>%
  mutate(upgraded = ifelse(book_class < travel_class, 1, 0)) # 1 indicates true
```

Create cardholder + uflystatus col

```r
data <- data %>%
  mutate(sc_rewards = ifelse(UflyMemberStatus == 'non-ufly', 1,
                      ifelse(UflyMemberStatus == 'Standard' & CardHolder == 'false', 2,
                      ifelse(UflyMemberStatus == 'Standard' & CardHolder == 'true', 3,
                      ifelse(UflyMemberStatus == 'Elite' & CardHolder == 'false', 4, 5)))))
```

Create dummy gender code col

```r
data <- data %>%
  mutate(gender = ifelse(GenderCode == 'F', 0, 1))
```

Create numerical col for booking channel, assuming reservations are less advantageous than Sun Country websites

```r
data <- data %>%
  mutate(booking_channel = ifelse(BookingChannel == 'Other', 1,
                           ifelse(BookingChannel == 'Outside Booking', 2,
                           ifelse(BookingChannel == 'Tour Operator Portal', 3,
                           ifelse(BookingChannel == 'Reservations Booking', 4,
                           ifelse(BookingChannel == 'SCA Website Booking', 5, 6))))))
```

Save current state of data for exploratory analysis later on

```r
trips <- data %>% filter(uid %in% sample_Uids)
```

```r
fare_oddities <- data %>% filter(BaseFareAmt > TotalDocAmt)
fare_normal <- data %>% filter(BaseFareAmt < TotalDocAmt )
base_zero <- data %>% filter(BaseFareAmt == 0.00)
eq <- data %>% filter(BaseFareAmt == TotalDocAmt)
```

Create reliable fare col based upon the dirtiness of the BaseFareAmt col identified from the code snip above

```r
data_fare <- data %>% # losing about 10% of data due to filters
  filter(BaseFareAmt != 0.00) %>%
  filter(BaseFareAmt <= TotalDocAmt)
```

Create clean dfs for attaching cluster assignments after clustering is done. Creating two because a substantial amount of rows are lost when including the clean fare col

```r
data <- data %>%
  select(-TicketNum, -PostalCode, -MarketingFlightNbr,
         -MarketingAirlineCode, -error, GenderCode)
data_fare <- data_fare %>%
  select(-TicketNum, -PostalCode, -MarketingFlightNbr,
         -MarketingAirlineCode, -error, -GenderCode)
```

Check for missing data. No relevant data missing

Grouping on PNR and customer uid. Also, creating an interval between booking date and service date of the first flight for a trip

```r
pnr_level_df <- data %>%
  group_by(PNRLocatorID, uid) %>%
```

```
  summarize(age = first(Age),
            gender = first(gender),
            member_status = first(sc_rewards),
            booking_channel = first(booking_channel),
            booking_product = first(BookedProduct),
            days_between_service = as.numeric(as.Date(strptime(min(ServiceStartDate),
                                                               "%Y-%m-%d")) -
                                              as.Date(strptime(first(PNRCreateDate),
                                                               "%Y-%m-%d"))),
            travel_class = mean(travel_class),
            upgrade = mean(upgraded))

pnr_level_df_fare <- data_fare %>%
  group_by(PNRLocatorID, uid) %>%
  summarize(age = first(Age),
            gender = first(gender),
            member_status = first(sc_rewards),
            booking_channel = first(booking_channel),
            booking_product = first(BookedProduct),
            days_between_service = as.numeric(as.Date(strptime(min(ServiceStartDate), "%Y-%m-%d")) -
                                              as.Date(strptime(first(PNRCreateDate), "%Y-%m-%d"))),
            travel_class = mean(travel_class),
            upgrade = mean(upgraded),
            fare = first(TotalDocAmt))
```

Create fellow travelers col for both dfs

```
i <- 2
pnr_level_df$PNRLocatorID <- as.character(pnr_level_df$PNRLocatorID)
pnr_level_df$group_size <- 1
for (id in pnr_level_df$PNRLocatorID){
  ifelse(id == as.character(pnr_level_df[i, 1]),
         pnr_level_df[i, 11] <- as.integer(pnr_level_df[i - 1, 11]) + 1,
         pnr_level_df[i, 11] <- as.integer(pnr_level_df[i, 11] + 0))
  i <- i + 1
}

max_group_sizes <- pnr_level_df %>%
  group_by(PNRLocatorID) %>%
  summarize(s = max(group_size))

pnr_level_df <- merge(pnr_level_df, max_group_sizes, by = 'PNRLocatorID')

i <- 2
pnr_level_df_fare$PNRLocatorID <- as.character(pnr_level_df_fare$PNRLocatorID)
pnr_level_df_fare$group_size <- 1
for (id in pnr_level_df_fare$PNRLocatorID){
  ifelse(id == as.character(pnr_level_df_fare[i, 1]),
         pnr_level_df_fare[i, 12] <- as.integer(pnr_level_df_fare[i - 1, 12]) + 1,
         pnr_level_df_fare[i, 12] <- as.integer(pnr_level_df_fare[i, 12] + 0))
  i <- i + 1
}

max_group_sizes_fare <- pnr_level_df_fare %>%
```

```
  group_by(PNRLocatorID) %>%
  summarize(s = max(group_size))

pnr_level_df_fare <- merge(pnr_level_df_fare, max_group_sizes_fare, by = 'PNRLocatorID')
```

Add a fare_per_traveler col for each PNR. This captures the relative profitability of each customer per flight

```
pnr_level_df_fare <- pnr_level_df_fare %>%
  mutate(fare_per_traveler = fare / s)
```

Save current state of dfs for exploratory analysis later on

```
pnrs <- pnr_level_df
pnrs_with_fare <- pnr_level_df_fare
```

Grouping on customer uid and create a total trips col

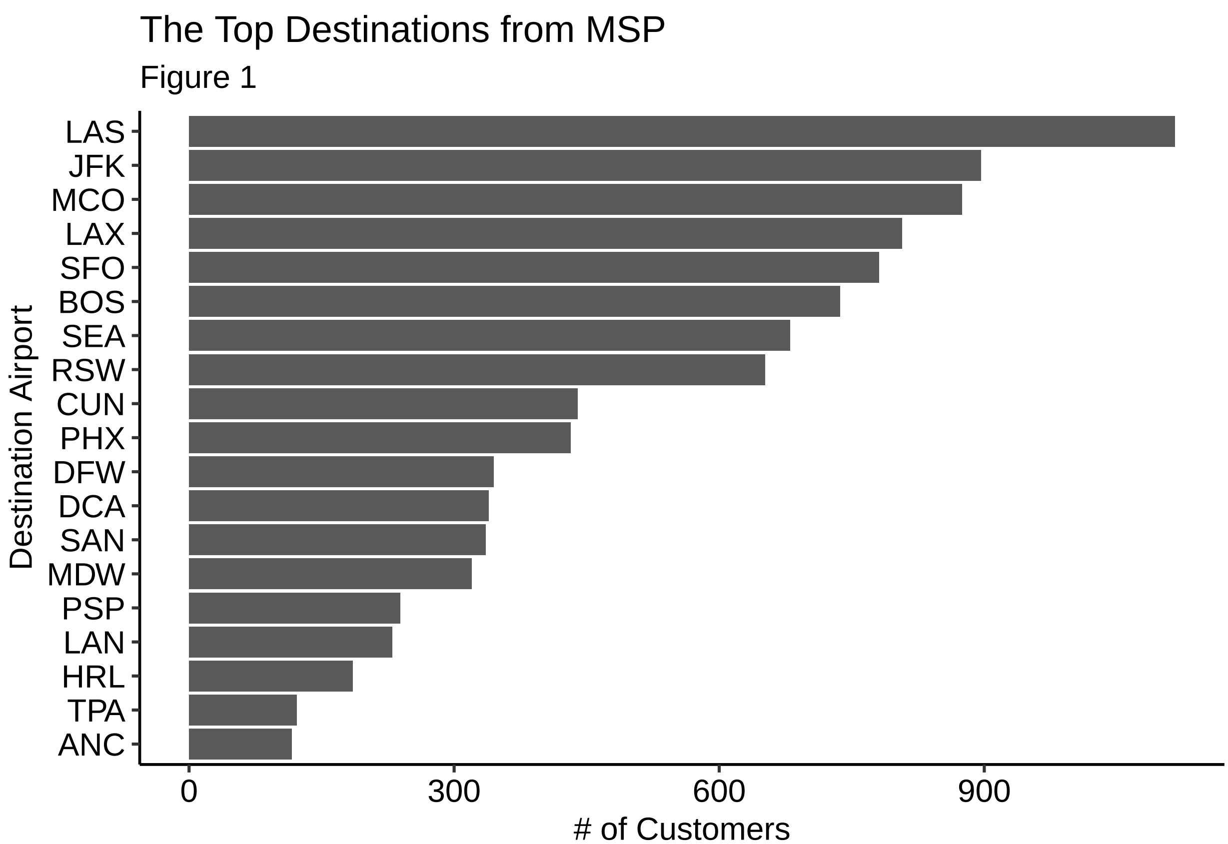
```
customers <- pnr_level_df %>%
  group_by(uid) %>%
  summarise(age = mean(age),
            gender = first(gender),
            member_status = mean(member_status),
            booking_channel = mean(booking_channel),
            days_between_service = mean(days_between_service),
            travel_class = mean(travel_class),
            upgrade_rate = mean(upgrade),
            fellow_travelers = mean(s),
            total_trips = n()) %>%
  arrange(-total_trips)

customers_fare <- pnr_level_df_fare %>%
  group_by(uid) %>%
  summarise(age = mean(age),
            gender = first(gender),
            member_status = mean(member_status),
            booking_channel = mean(booking_channel),
            days_between_service = mean(days_between_service),
            travel_class = mean(travel_class),
            upgrade_rate = mean(upgrade),
            fellow_travelers = mean(s),
            total_trips = n(),
            fare_per_traveler = mean(fare_per_traveler)) %>%
  arrange(-total_trips)
```

With the necessary dataframes created, the team could now answer the question at hand.

0.2.1 What are the top destinations SCA's customers fly to and from?

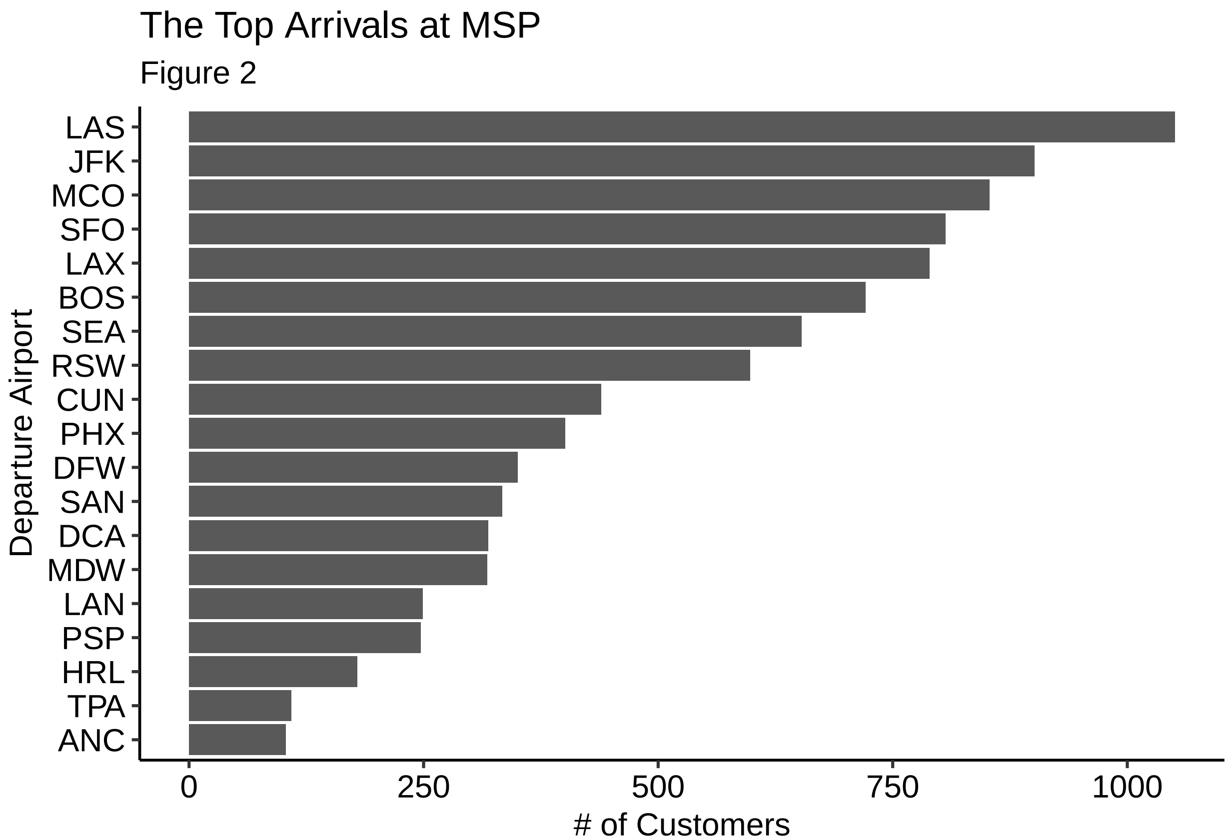
```
trips$ServiceEndCity <- as.character(trips$ServiceEndCity)
trips$ServiceStartCity <- as.character(trips$ServiceStartCity)

flights <- trips %>%
  filter(CouponSeqNbr <= 2) %>% # only losing a small # of flights
  select(ServiceStartCity, ServiceEndCity, ServiceStartDate, month, Season) %>%
  mutate(type = ifelse(ServiceStartCity == 'MSP', 'Departure', 'Arrival'))  %>%
  mutate(airport = ifelse(type == 'Departure', ServiceEndCity, ServiceStartCity))
```

```r
msp_out <- flights %>%
  filter(ServiceStartCity == 'MSP')

destinations <- msp_out %>%
  group_by(ServiceEndCity) %>%
  summarize(count = n()) %>%
  filter(count > 100) %>%
  arrange(-count)

destinations$ServiceEndCity <- as.factor(destinations$ServiceEndCity)
destinations$ServiceEndCity <- factor(destinations$ServiceEndCity,
      levels = destinations$ServiceEndCity[order(destinations$count)])

ggplot(destinations) + geom_col(aes(x = ServiceEndCity, y = count)) + coord_flip() +
  labs(title = 'The Top Destinations from MSP', subtitle = 'Figure 1',
       y = '# of Customers', x = 'Destination Airport') +
  theme_pubr()
```



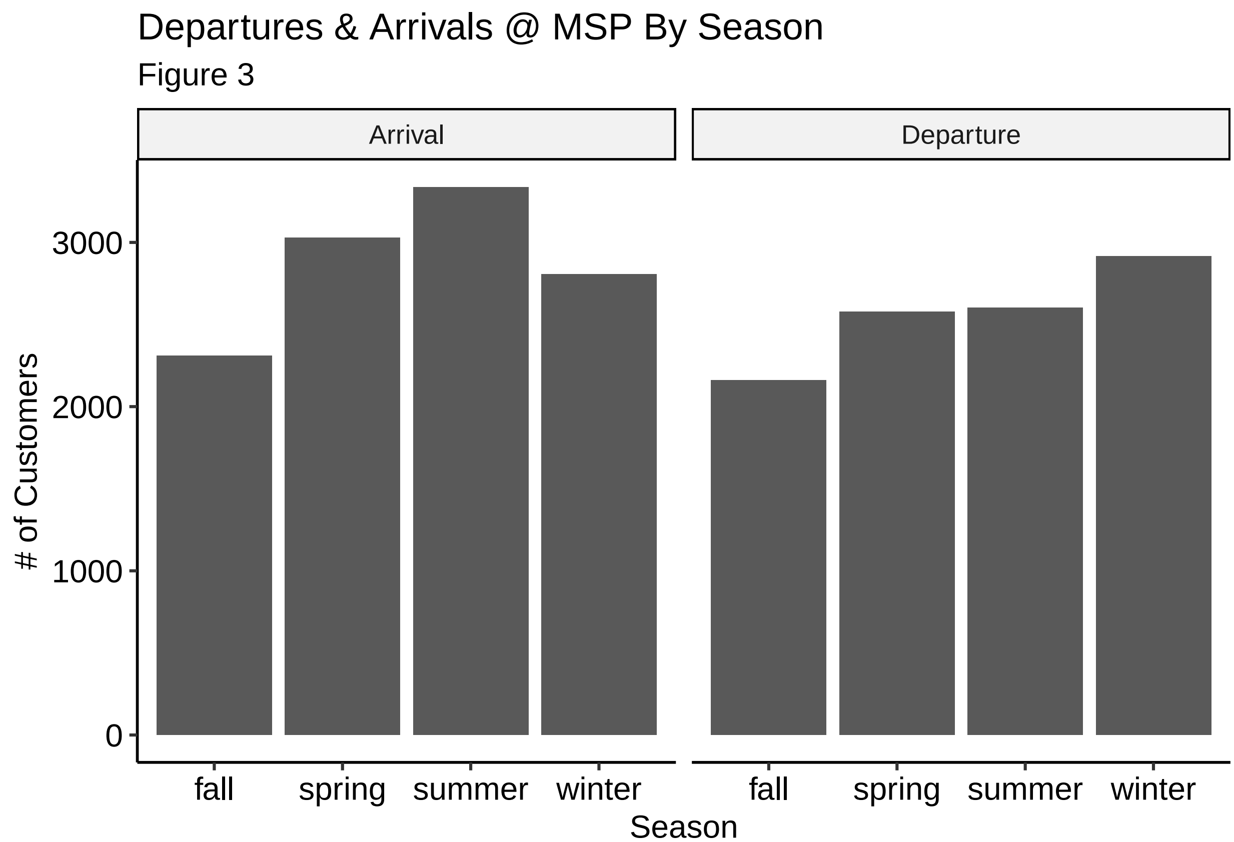

```r
msp_in <- flights %>%
  filter(ServiceEndCity == 'MSP')

departures <- msp_in %>%
  group_by(ServiceStartCity) %>%
  summarize(count = n()) %>%
  filter(count > 100) %>%
  arrange(-count)
```

```r
departures$ServiceStartCity <- as.factor(departures$ServiceStartCity)
departures$ServiceStartCity <- factor(departures$ServiceStartCity,
                levels = departures$ServiceStartCity[order(departures$count)])

ggplot(departures) + geom_col(aes(x = ServiceStartCity, y = count)) + coord_flip() +
  labs(title = 'The Top Arrivals at MSP', subtitle = 'Figure 2',
       x = 'Departure Airport', y = '# of Customers') + theme_pubr()
```



Interpretation & Conclusions Figures 1 & 2 were not very insightful, but they did gather a baseline understanding of SCA's flight offerings. The team hypothesized that it may be more informative to look at whether there are any flight patterns associated with seasonality.

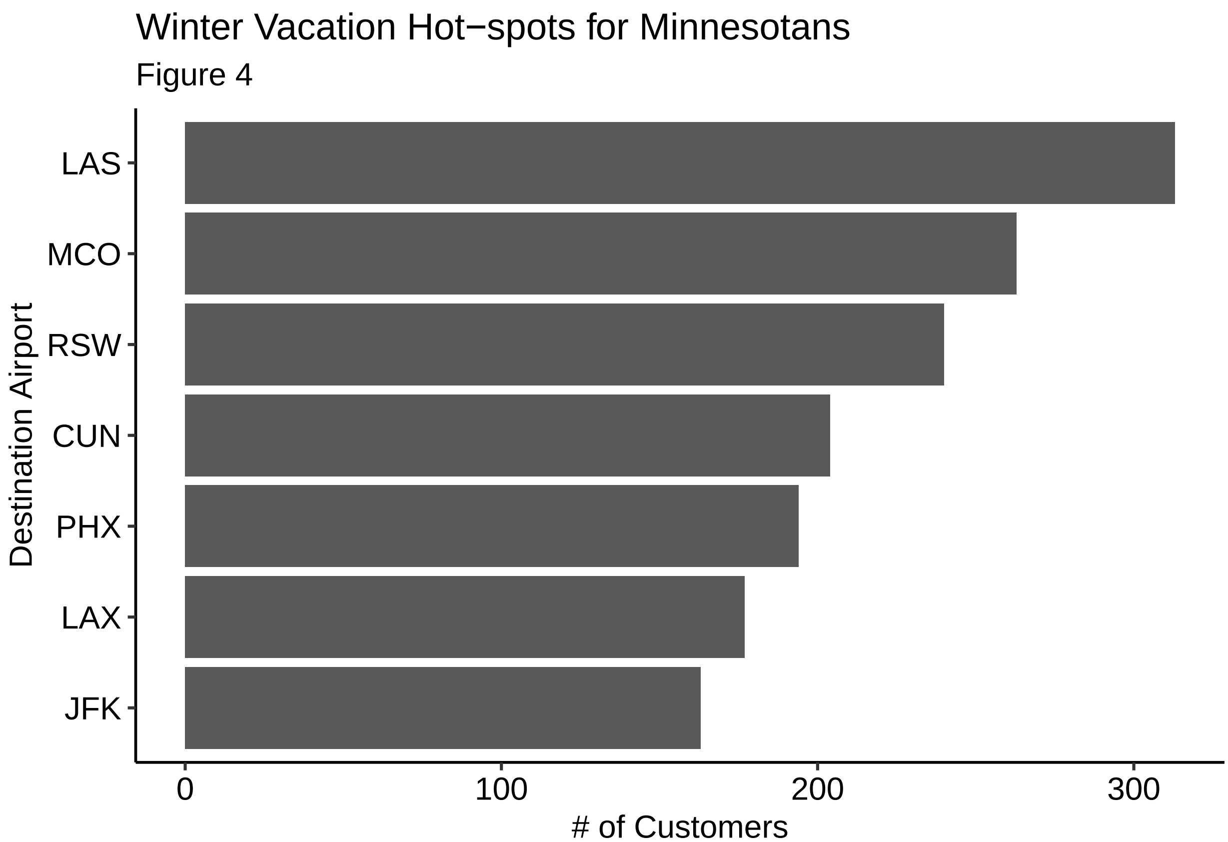
```r
flights$type <- as.factor(flights$type)

flights %>%
  group_by(type, Season) %>%
  summarize(count = n()) %>%
  ggplot() + geom_col(aes(x = Season, y = count)) + facet_grid(. ~ type) +
  labs(title = 'Departures & Arrivals @ MSP By Season',
       subtitle = 'Figure 3', y = '# of Customers') + theme_pubr()
```



Interpretation & Conclusions The volume of departures at MSP spikes during the winter and the volume of arrivals at MSP spikes during the summer.

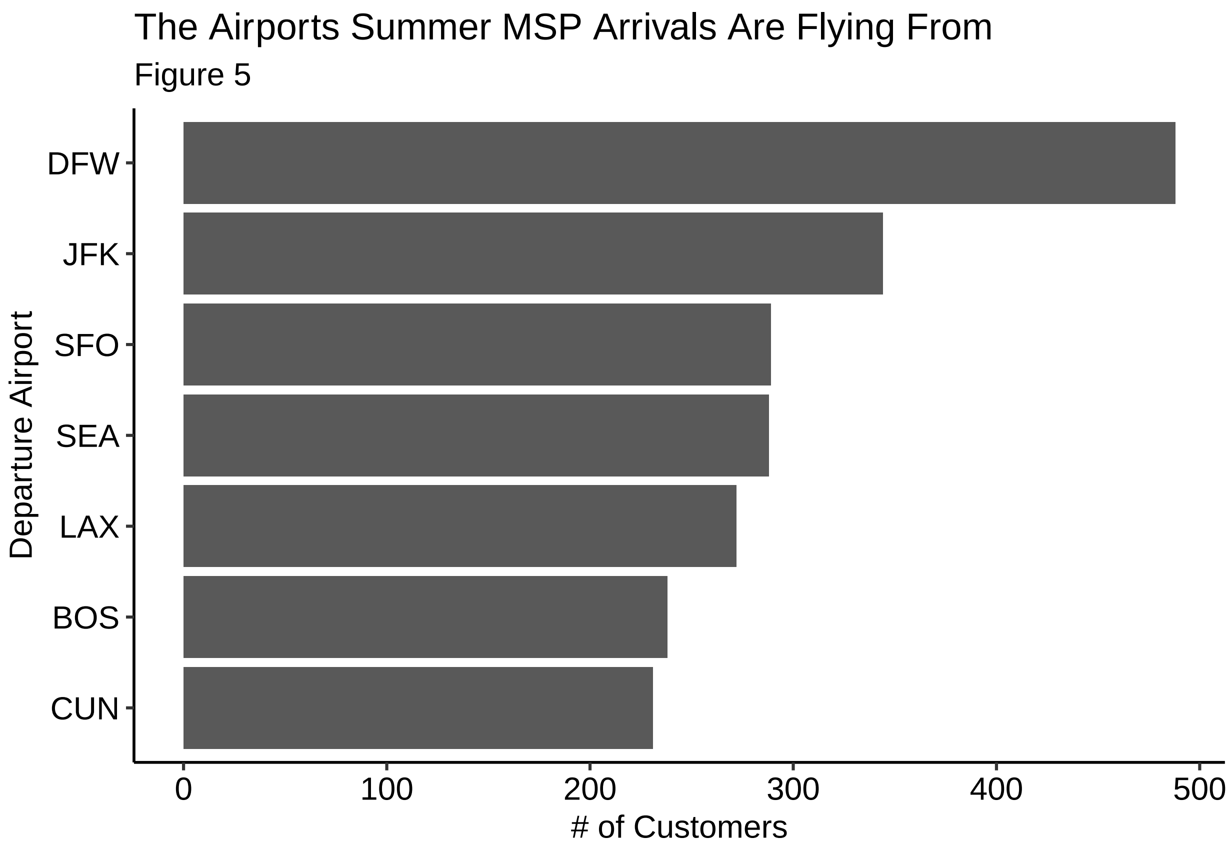
0.2.2 Where are Minnesotans vacationing in the winter and tourists coming from in the summer?

```
flights$airport <- as.factor(flights$airport)

z1 <- flights %>%
  filter(type == 'Departure' & Season == 'winter') %>%
  group_by(airport) %>%
  summarise(count = n()) %>%
  filter(count > 150)

z1$airport <- factor(z1$airport, levels = z1$airport[order(z1$count)])
ggplot(z1) + geom_col(aes(x = airport, y = count)) + coord_flip() +
  labs(title = 'Winter Vacation Hot-spots for Minnesotans', subtitle = 'Figure 4',
       x = 'Destination Airport', y = '# of Customers') +
  theme_pubr()
```



Interpretation & Conclusions Figure 4 confirmed the team's intuition that a lot of Minnesotans vacation in the winter in warmer places such as Las Vegas (LAS), Florida (MCO, RSW), Arizona (PHX), and Cancun (CUN) - Florida being the most popular destination when you combine the flights to MCO & RSW.

```
z2 <- flights %>%
  filter(type == 'Arrival' & Season == 'summer') %>%
  group_by(airport) %>%
  summarise(count = n())  %>%
  filter(count > 200)

z2$airport <- factor(z2$airport, levels = z2$airport[order(z2$count)])
ggplot(z2) + geom_col(aes(x = airport, y =count)) + coord_flip() +
  labs(title = 'The Airports Summer MSP Arrivals Are Flying From', subtitle = 'Figure 5',
       y = '# of Customers', x = 'Departure Airport') +
  theme_pubr()
```

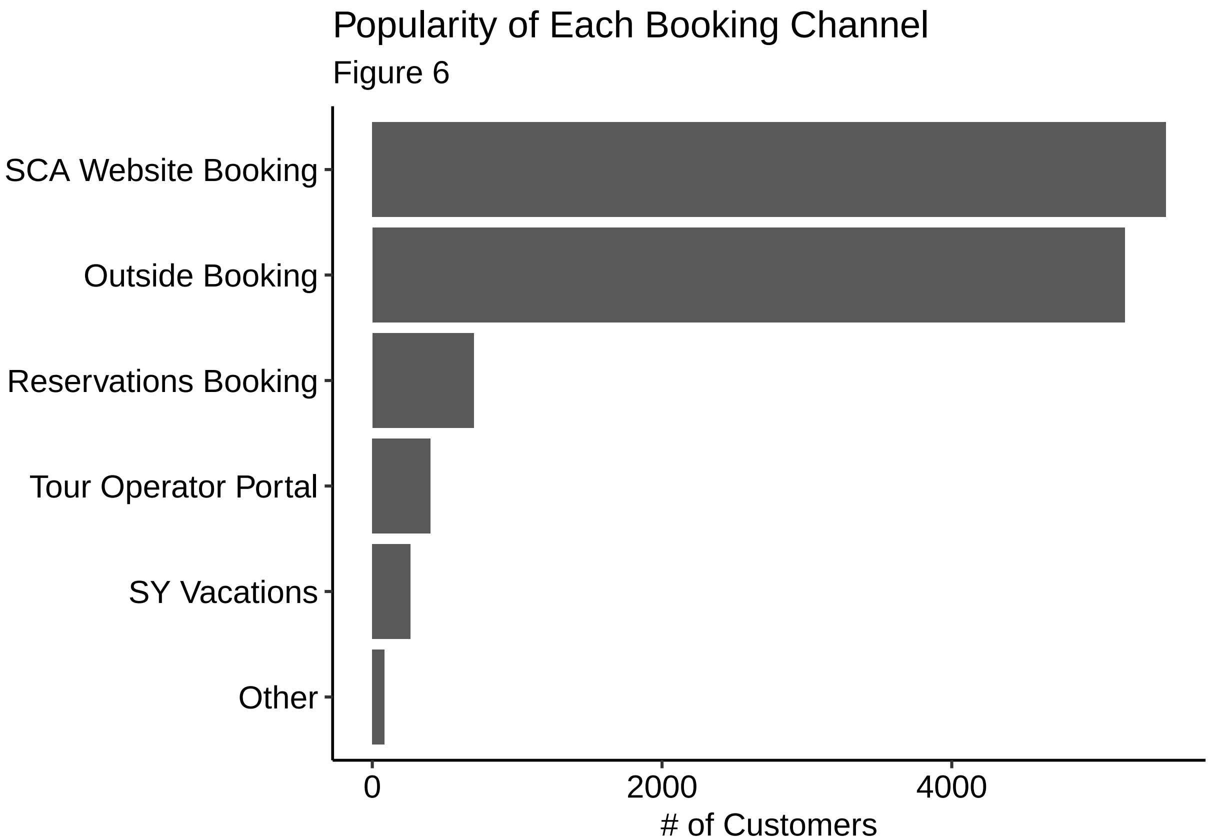


Interpretation & Conclusions Figure 5 raises an interesting insight. By far the most popular flight to MSP in the summer is from Dallas (DFW). As shown in figure 4, Dallas is not a popular winter vacation destination for Minnesotans. Thus, these travelers cannot be returning home from vacation. It is safe to assume from these two observations that SCA's customers from Dallas love the Minnesotan summers. This is a customer insight that SCA could forward to their tour agency partners to develop vacation packages around and utilize to target customers for promotional booking products.

0.2.3 What is the distribution of booking channel popularity for customers and profitability for SCA?

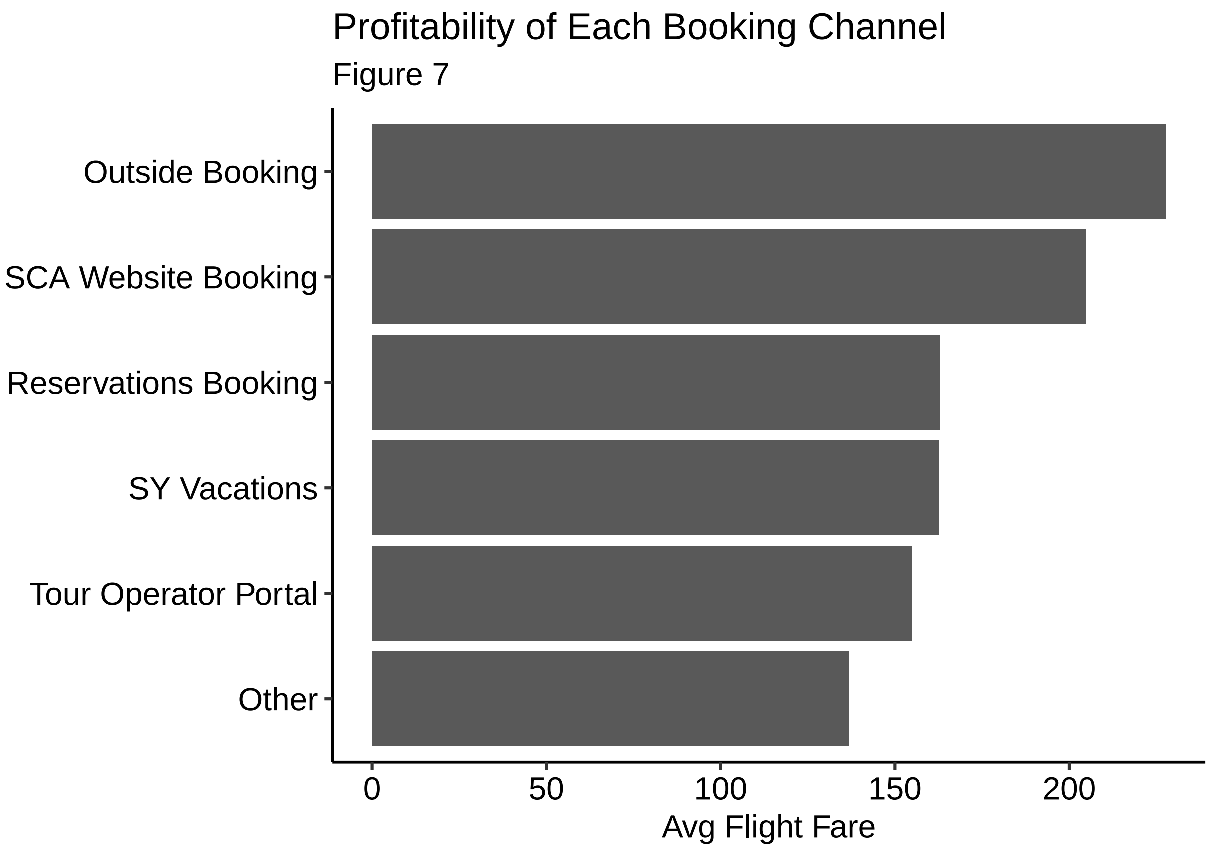
```
normal_booking_channel <- pnrs %>%  filter(uid %in% sample_Uids) %>%
  group_by(booking_channel) %>%
  summarize(count = n()) %>%
  arrange(-count)

normal_booking_channel <- normal_booking_channel %>%
  mutate(booking_channel_name = ifelse(booking_channel == 1, 'Other',
                          ifelse(booking_channel == 2, 'Outside Booking',
                          ifelse(booking_channel == 3, 'Tour Operator Portal',
                          ifelse(booking_channel == 4, 'Reservations Booking',
                          ifelse(booking_channel == 5, 'SCA Website Booking',
                                 'SY Vacations'))))))

normal_booking_channel$booking_channel_name <- factor(normal_booking_channel$booking_channel_name,
 levels = normal_booking_channel$booking_channel_name[order(normal_booking_channel$count)])

ggplot(normal_booking_channel) + geom_col(aes(x = booking_channel_name, y = count)) +
```

```r
  coord_flip() + labs(title = 'Popularity of Each Booking Channel',
  subtitle = 'Figure 6', y = '# of Customers', x = element_blank()) + theme_pubr()
```



Note, there is no visibility into the revenue Sun Country generates from flights booked through Tour Operator Portals. Thus, this channel is excluded from the plot below. The team is also assuming that the average flight fare for a customer is a reasonable metric to assess profitability of that seat

```r
channels <- pnrs_with_fare %>% filter(uid %in% sample_Uids) %>%
  select(booking_channel, fare_per_traveler) %>%
  group_by(booking_channel) %>%
  summarize(mean_fare = mean(fare_per_traveler)) %>%
  arrange(-mean_fare)

channels <- channels %>%
  mutate(booking_channel_name = ifelse(booking_channel == 1, 'Other',
                         ifelse(booking_channel == 2, 'Outside Booking',
                         ifelse(booking_channel == 3, 'Tour Operator Portal',
                         ifelse(booking_channel == 4, 'Reservations Booking',
                         ifelse(booking_channel == 5, 'SCA Website Booking',
                                'SY Vacations'))))))

channels$booking_channel_name <- factor(channels$booking_channel_name,
                levels = channels$booking_channel_name[order(channels$mean_fare)])
ggplot(channels) + geom_col(aes(x = booking_channel_name, y = mean_fare)) +
  coord_flip() + labs(title = 'Profitability of Each Booking Channel',
  subtitle = 'Figure 7', x = element_blank(), y = 'Avg Flight Fare') + theme_pubr()
```



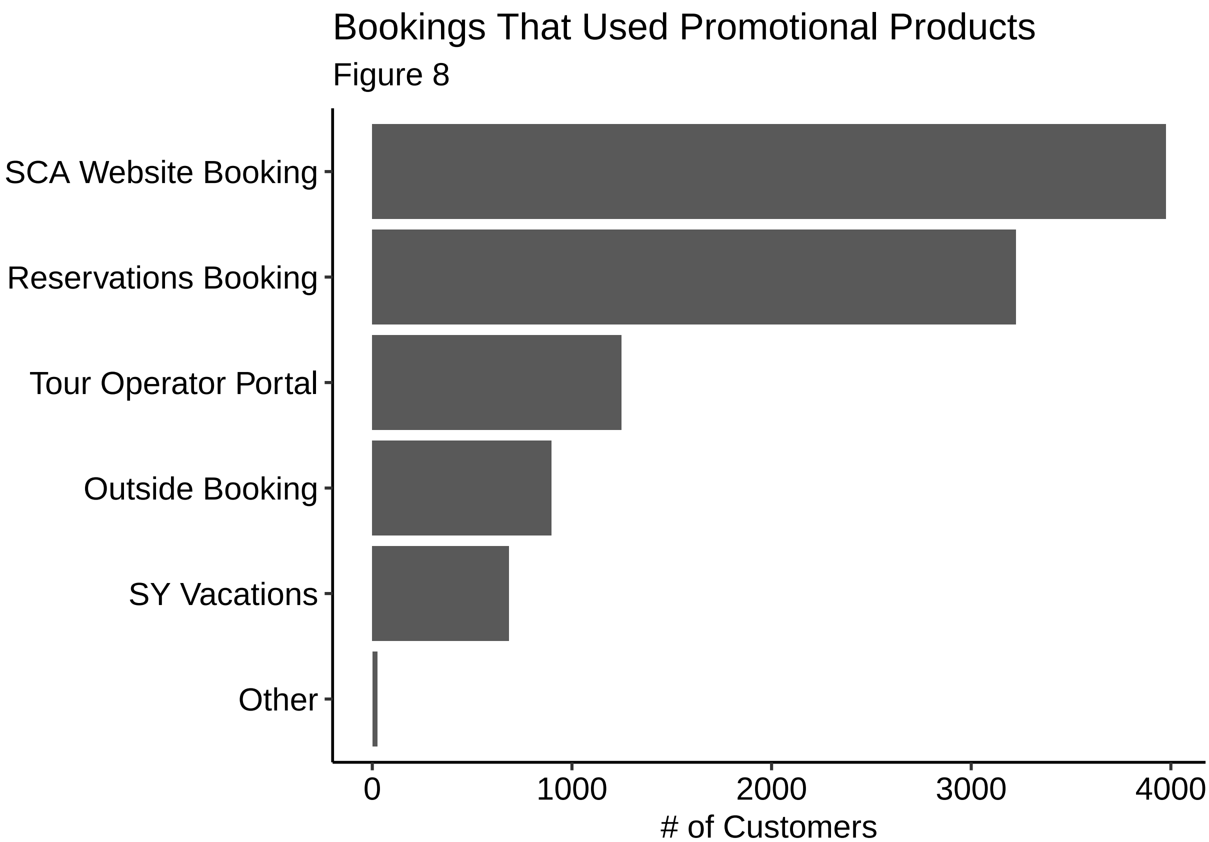
Interpretation & Conclusions The bookings from SCA's general website brought in a substantial increase in revenue per seat compared to bookings through SY Vacations. It is unclear as to why SY Vacations is not utilized by customers or profitable for SCA. Nonetheless, figures 7 & 8 raise concerns for the SY Vacations booking channel.

0.2.4 Out of the people who do use SCA's promotional booking products, what channel do they use to book flights?

```
top_booking_products <- pnrs %>% filter(uid %in% sample_Uids) %>%
  filter(booking_product != '') %>%
  group_by(booking_product) %>%
  summarise(count = n()) %>%
  arrange(-count) %>%
  filter(count > 100) %>% # Number of flights minimum to be a prominent booking product
  select(booking_product, count)

top_booking_products <- pnrs %>%
   filter(booking_product %in% top_booking_products$booking_product)
```

```
better_booking_channel <- top_booking_products %>%
  group_by(booking_channel) %>%
  summarize(count = n()) %>%
  arrange(-count)

better_booking_channel <- better_booking_channel %>%
  mutate(booking_channel_name = ifelse(booking_channel == 1, 'Other',
                         ifelse(booking_channel == 2, 'Outside Booking',
```

```
                                    ifelse(booking_channel == 3, 'Tour Operator Portal',
                                    ifelse(booking_channel == 4, 'Reservations Booking',
                                    ifelse(booking_channel == 5, 'SCA Website Booking',
                                           'SY Vacations'))))))

better_booking_channel$booking_channel_name <- factor(better_booking_channel$booking_channel_name,
    levels = better_booking_channel$booking_channel_name[order(better_booking_channel$count)])
ggplot(better_booking_channel) + geom_col(aes(x = booking_channel_name, y = count)) +
  coord_flip() + labs(title = 'Bookings That Used Promotional Products',
  subtitle = 'Figure 8', x = element_blank(), y = '# of Customers') + theme_pubr()
```

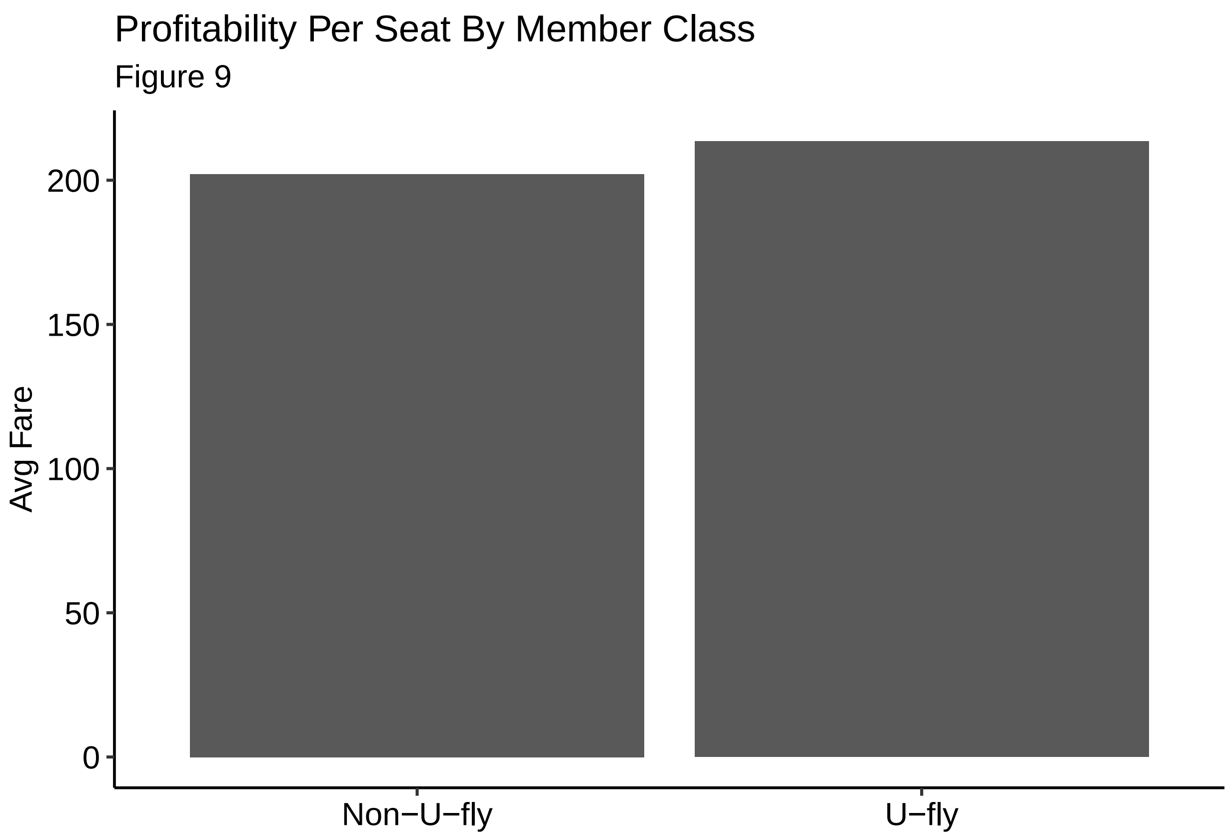


Interpretation & Conclusions Only 35% of flights are booked using one of Sun Country's prominent booking products. Depending upon the strategy and investment of SCA's marketing team, this statistic could be considerably low. After looking into which booking channel promotional products are consistently used with, Tour Operator Portals seemed to be pushing Sun Country's prominent promotional products relatively well after seeing this channel jump to third in popularity in figure 8 from fifth in popularity in figure 7. How can SCA facilitate more bookings through the Tour Operator Portal channel to make better use of their promotional booking products?

0.2.5 Are there demonstrable differences in profitability and customer loyalty between U-fly members and non-U-fly members?

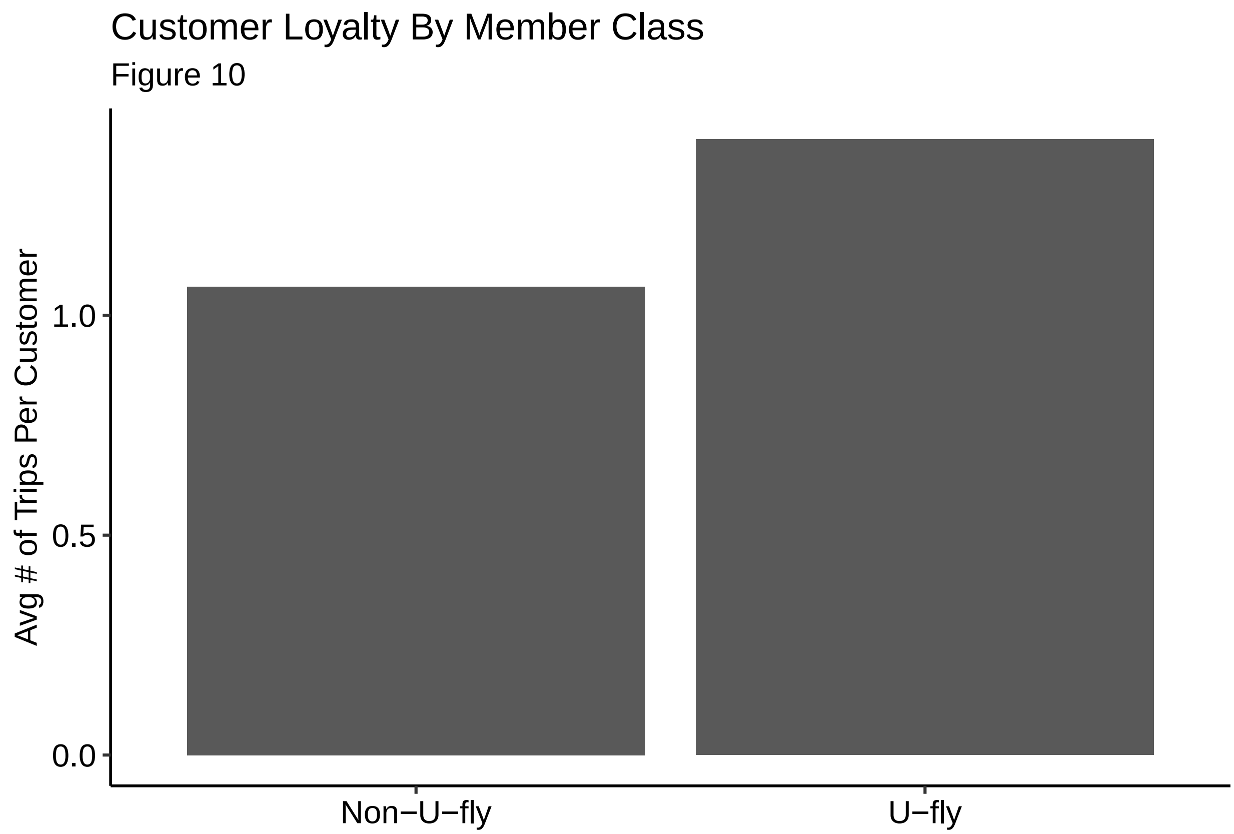
```
customers <- customers %>%
  mutate(Member_Status = ifelse(member_status == 1, 1, 2))

customers_fare <- customers_fare %>%
  mutate(Member_Status = ifelse(member_status == 1, 1, 2))
```

```r
customers_fare %>% filter(uid %in% sample_Uids) %>%
  filter(member_status %in% c(1, 2, 3, 4, 5)) %>%
  filter(fare_per_traveler < 1000) %>% # more accurate representation of the typical customer
  mutate(Member_Status_name = ifelse(Member_Status == 1, 'Non-U-fly', 'U-fly')) %>%
  group_by(Member_Status_name) %>%
  summarize(mean_fare = mean(fare_per_traveler)) %>%
  ggplot() + geom_col(aes(x = as.factor(Member_Status_name), y = mean_fare)) +
  labs(title = 'Profitability Per Seat By Member Class', subtitle = 'Figure 9',
       y = 'Avg Fare', x = element_blank()) + theme_pubr()
```



Below the team is assuming the total # of trips a customer has booked with SCA is a reasonable metric to assess customer loyalty

```r
customers %>%
  mutate(Member_Status_name = ifelse(Member_Status == 1, 'Non-U-fly', 'U-fly')) %>%
  group_by(Member_Status_name) %>%
  summarize(mean_trips = mean(total_trips)) %>%
  ggplot() + geom_col(aes(x = as.factor(Member_Status_name), y = mean_trips)) +
  labs(title = 'Customer Loyalty By Member Class', subtitle = 'Figure 10',
       y = 'Avg # of Trips Per Customer', x = element_blank()) + theme_pubr()
```



Interpretation & Conclusions Sun Country improves profitability per seat and customer loyalty when customers join U-Fly. The program is accomplishing what it set out to do.

0.2.6 Are there certain customers who are susceptible to travel class upgrades?

```
upgraders <- customers %>% filter(uid %in% sample_Uids) %>%
  filter(upgrade_rate > 0.4) %>%
  summarise(avg_age = mean(age),
            avg_member_status = mean(Member_Status),
            avg_booking_channel = mean(booking_channel),
            avg_fellow_travelers = mean(fellow_travelers))

non_upgraders <- customers %>% filter(uid %in% sample_Uids) %>%
  filter(upgrade_rate < 0.4) %>%
  summarise(avg_age = mean(age),
            avg_member_status = mean(Member_Status),
            avg_booking_channel = mean(booking_channel),
            avg_fellow_travelers = mean(fellow_travelers))

u <- rbind(upgraders, non_upgraders)
u['customer_type'] <- c('Upgraders', 'Non-upgraders')
u <- u %>%
  select(customer_type, avg_age, avg_fellow_travelers, avg_member_status)
knitr::kable(u)
```

customer_type	avg_age	avg_fellow_travelers	avg_member_status
Upgraders	46.08871	1.892890	1.326568

customer_type	avg_age	avg_fellow_travelers	avg_member_status
Non-upgraders	39.02868	2.570389	1.164746

Interpretation & Conclusions There is indeed a specific customer profile that upgrades more often. Upgraders typically are older, fly with fewer people, and are more often U-fly members.

The two key takeaways from the initial customer behavioral analysis:

1. SCA can target two distinct customer sets for vacation packages:Dallas to MSP summer travelers & MSP to Florida winter travelers. The team has termed these customer groups as "Minnesota Vacationers" & "Florida Vacationers".
2. SCA obtains more revenue per seat and a higher level of customer loyalty when customers are U-fly members. SCA should identify susceptible customers to target that have a higher probability of joining the U-fly program than all customers on average. Clustering has been able to assist in this targeting effort as shown in the next section. The team hypothesized that the customers identified as susceptible to joining U-fly from the clustering analysis will have a similar customer profile as the "Upgraders" customer profile shown above.

0.3 Are there distinct groups of customers that fly with Sun Country?

Execution and Results:

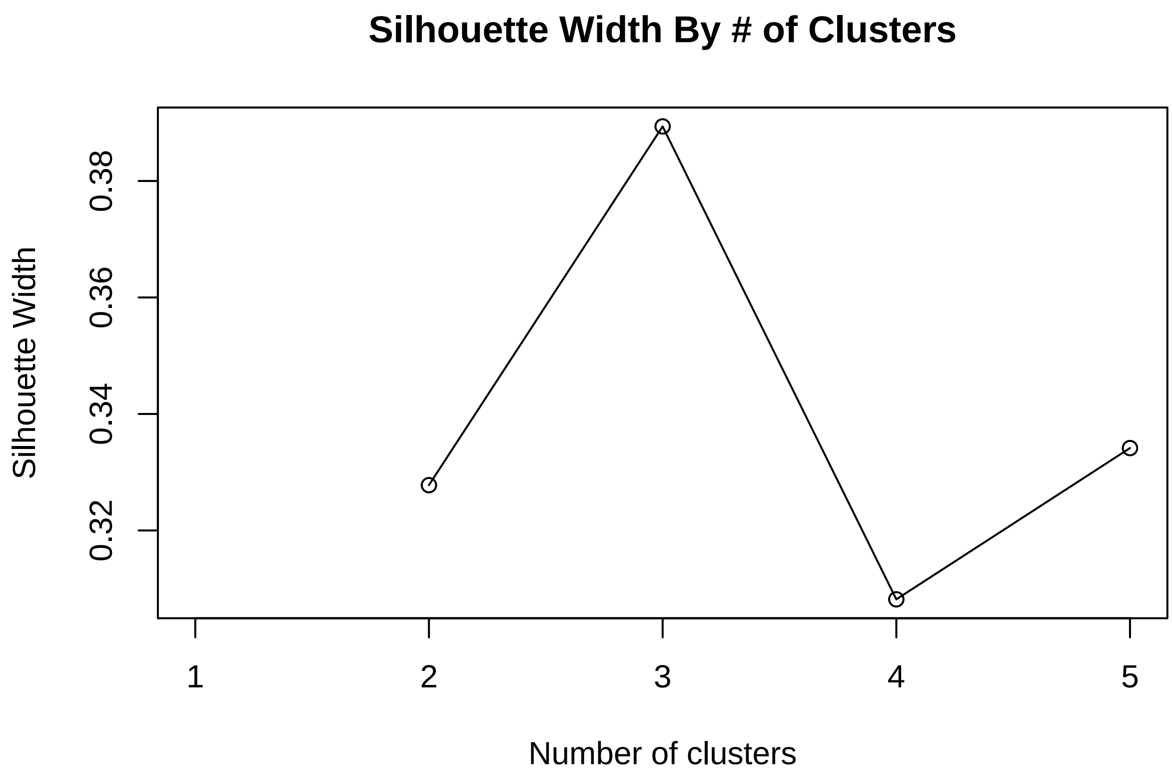
```
normalize <- function(x){
  return ((x - min(x))/(max(x) - min(x)))
}
```

```
clust_df2 <- customers_fare %>% filter(uid %in% sample_Uids) %>%
  select(uid:fare_per_traveler)

clust_df2$ufly <- 'non-ufly'
clust_df2$ufly[which(round(clust_df2$member_status)>=2 )] <- 'ufly'
clust_df2$ufly <- as.factor(clust_df2$ufly)


clust_df2_std <- clust_df2 %>% select(age, days_between_service:fare_per_traveler)%>%
  mutate(age = normalize(age),
            upgrade_rate = upgrade_rate,
            travel_class = normalize(travel_class),
            group_size = normalize(fellow_travelers),
            days_between_service = normalize(days_between_service),
            total_trips = normalize(total_trips),
            fare_per_traveler = normalize(fare_per_traveler))

gower_dist <- daisy(clust_df2_std, metric = "gower")
```

Calculate silhouette width using PAM

```
sil_width_1 <- c(NA)
for(i in 2:5){
  pam_fit_1 <- pam(gower_dist,
                  diss = TRUE,
                  k = i)
  sil_width_1[i] <- pam_fit_1$silinfo$avg.width
}
```

```r
plot(1:5, sil_width_1,
     main = 'Silhouette Width By # of Clusters',
     xlab = "Number of clusters",
     ylab = "Silhouette Width")
lines(1:5, sil_width_1)
```



Interpretation & Conclusions Three clusters were the best choice since we are trying to get the silhouette width as close to one as possible. When the silhouette width is close to one that indicates there is quality homogeneity within clusters and heterogeneity across clusters.

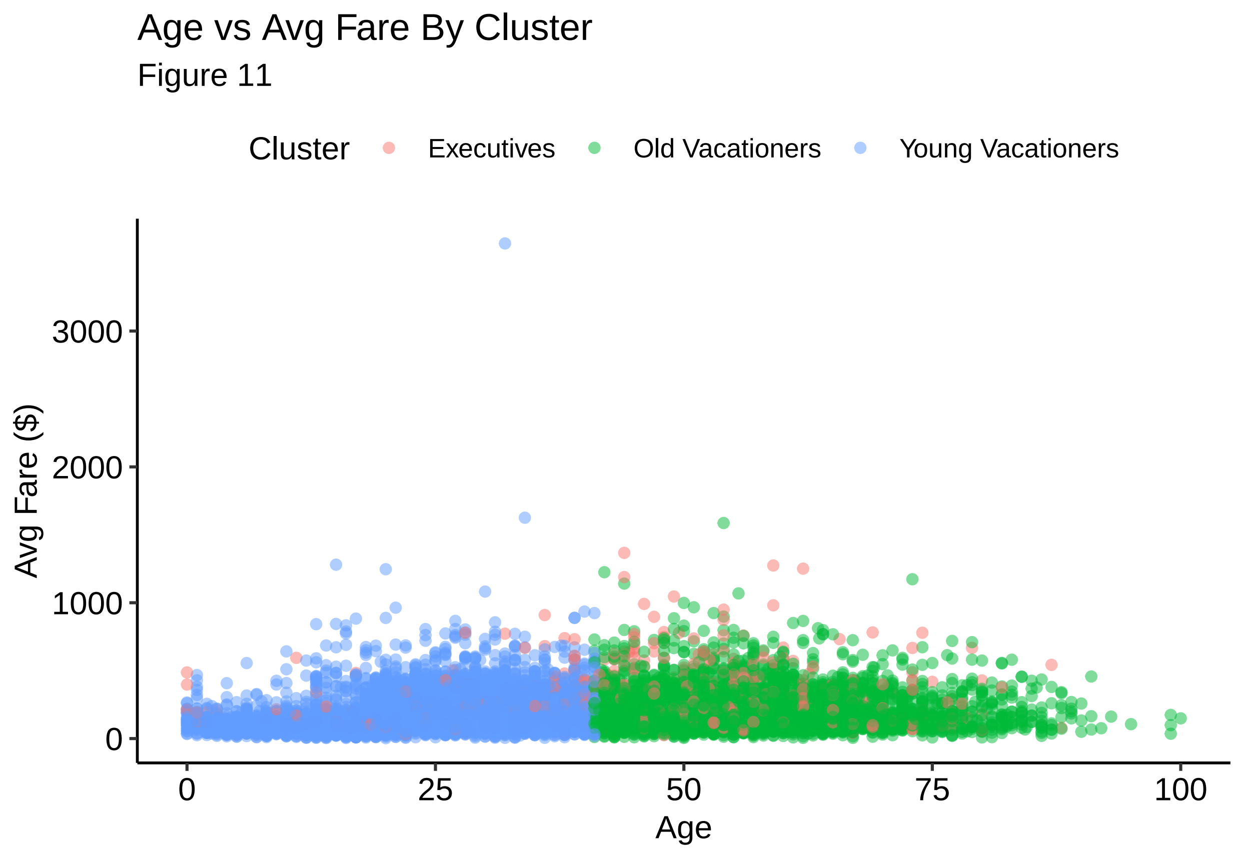
```r
pam_fit_3K <- pam(gower_dist, diss = TRUE, k = 3)
```

With the clusters generated, they must now be interpreted for business use. Below is an initial exploratory plot and the means table of pertinent customer attributes

```r
data1 <- pam_fit_3K$clustering %>%
  data.frame() %>%
  setNames('cluster') %>%
  mutate(uid = clust_df2$uid)

df_customer_cluster <- merge(data1, clust_df2, by.x = 'uid',
                             by.y = 'uid')
```

```r
df_customer_cluster$cluster <- as.factor(df_customer_cluster$cluster)
df_customer_cluster <- df_customer_cluster %>%
  mutate(Cluster = ifelse(cluster == 1, 'Old Vacationers',
                   ifelse(cluster == 2, 'Young Vacationers', 'Executives')))

ggplot(df_customer_cluster) + geom_point(aes(x = age, y=fare_per_traveler,
                                         color = Cluster),alpha = 0.5) +
  labs(title = 'Age vs Avg Fare By Cluster', x = 'Age',
```

```
  y = 'Avg Fare ($)', subtitle = 'Figure 11') + theme_pubr()
```



Recall the encoding for the attribute travel_class is: 1 = Coach, 2 = Discount First Class, 3 = First Class

```
pam_results_3K <- clust_df2 %>%
  mutate(cluster = pam_fit_3K$clustering) %>%
  select(-booking_channel, -gender, -days_between_service, -uid) %>%
  mutate(uflyc = ifelse(ufly == 'ufly', 1, 0)) %>%
  mutate(non_uflyc = ifelse(ufly == 'non-ufly', 1, 0)) %>%
    mutate(Cluster = ifelse(cluster == 1, 'Old Vacationers',
                    ifelse(cluster == 2, 'Young Vacationers', 'Executives'))) %>%
  group_by(Cluster) %>%
  summarize(customers = n(), age = mean(age), avg_fare = mean(fare_per_traveler),
            travel_class = mean(travel_class), fellow_travelers = mean(fellow_travelers),
            upgrade_rate = mean(upgrade_rate), `%_ufly` = sum(uflyc)/(sum(uflyc)+ sum(non_uflyc)),
            total_trips = mean(total_trips)) %>%
  select(customers, `%_ufly`, age, avg_fare,
         travel_class, fellow_travelers, upgrade_rate)

knitr::kable(pam_results_3K)
```

customers	%_ufly	age	avg_fare	travel_class	fellow_travelers	upgrade_rate
598	0.3511706	47.43618	293.5033	2.215766	1.773381	0.6181402
3690	0.2243902	57.23370	210.4445	1.043596	2.304607	0.0014436
4568	0.1280648	24.18313	191.7560	1.024502	2.732011	0.0006367

Interpretation & Conclusions From the k-Mediods clustering analysis, three distinct clusters have been identified. One cluster, termed "Old Vacationers", is made up of vacationers (fellow travelers > 2) who fly coach (travel_class = 1.04) that rarely upgrade or join u-fly (proportion_ufly = 22.4%). There is another very similar cluster termed "Young Vacationers". These customers tend to be in their early twenties instead of late fifties and purchase cheaper flights. The last cluster, "Executives", is the unique cluster having only 598 customers compared to the other two clusters which both contain 3500+ customers. Executives seems to be the affluent (fare_per_traveler = $293), business travelers (fellow_travelers < 2) who are more often U-fly members (proportion_ufly = 35.1) than the other clusters, always travel first class, and are susceptible to upgrades. Executives are indeed similar to the "Upgraders" customer profile.

0.4 How can Sun Country leverage the clusters of customers to increase adoption of U-Fly and sell upgrades?

Executives is clearly the cluster that has the money to spare for upgrades and indulges in the luxuries of U-fly membership and first-class traveling. However, 65% of the customers in the Executives cluster are not U-fly members yet. Targeting this 65% of the Executives cluster would produce a far greater U-fly adoption rate than targeting the entire SCA customer population.

Execution and Results:

```r
executives <- df_customer_cluster %>%
  filter(Cluster == 'Executives')

ufly_targets <- executives %>%
  filter(member_status == 1) %>%
  mutate(Gender = ifelse(gender == 1, 'Male', 'Female')) %>%
  select(uid, age, Gender)

ufly_targets <- merge(ufly_targets, data, by = 'uid')

ufly_targets <- ufly_targets %>%
  select(uid, PaxName, birthdateid, Age, Gender) %>%
  unique()
```

Interpretation & Conclusions The ufly_targets dataframe contains 422 customers' unique id, birthdateid, paxname, age, and gender who are not yet U-fly members but are very similar to the customers in the Executives cluster who are already U-fly members. Thus, this is the customer set that SCA should target to join U-fly.

0.5 Final Recommendations

Below are the two marketing strategies the Carlson Consulting Team has developed for Warnken and Vaughn. These strategies will provide a more intimate vacation experience for customers, improve the ROI of promotional booking products, and boost the adoption rate of the U-fly membership. Over time, these benefits will facilitate a substantial improvement to the overall bottom-line for SCA and assist in competitively differentiating SCA from the larger airlines.

1)

Minnesota Vacationers and Florida Vacationers are ideal customer sets to target for vacation packages. The team advises SCA to directly market to these customers to allow them the flexibility to book through their preferred channel and to motivate their tour agency partners to market to them as well. Tour agencies are effective in pushing SCA's promotional booking products and should be leveraged as an additional marketing resource more than they currently are. This will help SCA get the most out of their promotional booking products which are only being used for 35% of bookings currently.

Both customer sets clearly rely upon SCA for their yearly vacation needs. SCA would enhance the customer experience for the Florida Vacationers by partnering with the top hospitality providers in Orlando (Airport = MCO) such as Disney & Universal. These partnerships would allow SCA to offer vacation bundles to the Florida Vacationers. These bundles could include travel to & from the airport, discount hotel & amusement park rates, and hand-selected excursions. To leverage the Florida Vacationers who are vacationing in Fort Myers, FL (Airport = RSW), SCA should partner with the Minnesota Twins to offer a similar vacation bundle as mentioned above but collaborating with the Twins spring training program. A third vacation bundle should be marketed to Minnesota Vacationers that simplifies the vacation planning experience for these customers and facilitates a higher quality vacation. SCA likely has the resources to cultivate unique Minnesotan vacation experiences better than these customers from Dallas do themselves.

2)

The customers tagged in the ufly_targets dataframe are much more likely to sign up for U-fly than the average SCA customer. By targeting this customer list instead of all customers, SCA will improve the adoption rate of U-fly. This will be a major benefit for the airline as SCA won't waste money targeting unlikely adopters anymore and these customers won't have to deal with the nuisance of being marketed another item they don't want in this over-saturated marketing world. All customers in the Executives cluster are quality targets for upgrading to first-class and discount first-class seats that aren't filled as well. The customers in the Executives cluster who have already upgraded are likely to continue to upgrade and the customers in the Executives cluster who haven't upgraded before are more likely to upgrade in the future than the average SCA customer.