HW2 EDA

Zhanglin Shangguan

2022/2/10

Table of contents

This technical document will be covering the following topics

1.Business Problem 2.Data Cleaning/Usability of the Data 3.Data Analysis 4.Customer Segmentation Analysis 5.Booking channel analysis 6.Conclusion Recommendations

Business Problem and Approach

Background:

Sun Country is currently facing intense competition with large brands in the airline industry. To help compete Sun Country is using different strategies to increase their market share including sending targeted ads to their customers.

The problem:

Sun Country's current advertisement targeting is only based on anecdotal evidence of customer behaviors.

Goal and solution:

Our team's goal is to discover if there are any hidden characteristics of Sun Country's customers. To solve this our team will use a data driven approach to segment Sun County's customers into defined groups based on demographics.

Data Cleaning

To get a more representative results, we decided to remove unrealistic outliers and performed feature engineering. For unrealistic outliers, we filtered out records with age greater than 110 because such population are very less likely to fly due to physical limitations, adn we removed records when gender is unidentified because it's largely error, those records does not provide us useful insights when we are studying gender-specific traveling pattern.

We converted all date columns into date formate, which allows to us to calculate the number of days before joining the Ufly Membership and days in advance that customer booked their trip before traveling. To take account the seasonal traveling pattern, we created column to record the month and the day of the week of all flights.

Also, it is important to note that we only focus on Sun Country by setting MarketingAirlineCode to "SY"

As customer segmentation analysis generates key insights for our study, we want to perform feature enginering to understand how combined variables might be able to provide more insights through clustering. Here, we created a new variable "Status" by combining "UflyMemberStatus" and "CardHolder" and categories including "Non-Member", "Status", "Standard Cardholder", "Elite", "Elite Cardholder".

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

## Filtering out extreme age groups, ungendered groups
## The Airlines with which booking was made is SY
## Combining UflyMemberStatus and CardHolder together to create new column Status
df <- df %>%
  filter(Age > 0 & Age < 110 & MarketingAirlineCode == "SY" & GenderCode != "U") %>%
  unite("Status", UflyMemberStatus:CardHolder, remove = TRUE)

## Categorize Status into 5 groups
df$Status[which(df$Status == "_")] = "Non-Member"
df$Status[which(df$Status == "Standard_false")] = "Standard"
df$Status[which(df$Status == "Standard_true")] = "Standard Carholder"
df$Status[which(df$Status == "Elite_false")] = "Elite"
df$Status[which(df$Status == "Elite_true")] = "Elite Cardholder"


df <- df %>%
  mutate(EnrollDate = ymd_hms(EnrollDate), ServiceStartDate = ymd(ServiceStartDate),
         PNRCreateDate = ymd(PNRCreateDate)) %>% # Convert dates into date format
  mutate(EnrollDate = floor_date(EnrollDate, "day")) %>% #Convert datetime into date format
  # The number of days of memebership before the flight
  mutate(memebership_duration_before_flight = difftime(ServiceStartDate, EnrollDate, units = "days"),
         # How many days in advance that customer booked their trip before traveling
         Daysinadvance_booked_flight = difftime(ServiceStartDate, PNRCreateDate, units = "days")) %>%
  #The month and the day of the week of flight, 1 is Sunday for day_of_week_flight
  mutate(month_flight = month(ServiceStartDate), day_of_week_flight = wday(ServiceStartDate, abbr = FALS
  # If customer have discounts, 0 - no discount, 1 - disount
  mutate(discount = if_else(BookedProduct == "", 0, 1)) %>%
  # Creating Stoppover column to indicate the duration
  mutate(Stopover = case_when(
    StopoverCode == "" ~ "Direct Flight",
    StopoverCode == "O" ~ "Layovers",
    StopoverCode == "X" ~ "24hrs Stopover")) %>%
  # unique Id
  unite("Unique_ID", c(EncryptedName, birthdateid, TicketNum) ,remove = FALSE)
```

As the data does not have unique identifier, we created "Unique_ID" by combining "EncryptedName", "birthdateid", and "TicketNum", and this identifier allows us to study the traveling frequency on individual level and helps us group each customers into frequent flyer or Non-frequent flyer. We chose this threshold to be 3 becuase each round trip counts as 2 times of flying, anyone travel 3 times suggest that they are flying more than 1 trip with Sun Country.

```
#Frequqent Flyer

df1 = df %>%
  group_by(Unique_ID) %>%
  dplyr::summarise(countv = n())
```

```
df <- join(df, df1,type = "left", by ="Unique_ID")

## define frequent flyer
df <- df%>%  mutate(
      FrequentFlyer =
      case_when(
        countv >= 3 ~ "Frequent Flyer",
        TRUE ~ " Not Frequent Flyer")
      )
```

Last but not least, we converted all categorical variables into factors, so that we can easily use them as inputs into clustering.

```
df$GenderCode = factor(df$GenderCode)
df$TrvldClassOfService = factor(df$TrvldClassOfService)
df$BookingChannel = factor(df$BookingChannel)
df$BkdClassOfService = factor(df$BkdClassOfService)
df$Stopover = factor(df$Stopover)
df$Status = factor(df$Status)
df$month_flight = factor(df$month_flight)
df$FrequentFlyer = factor(df$FrequentFlyer)
```

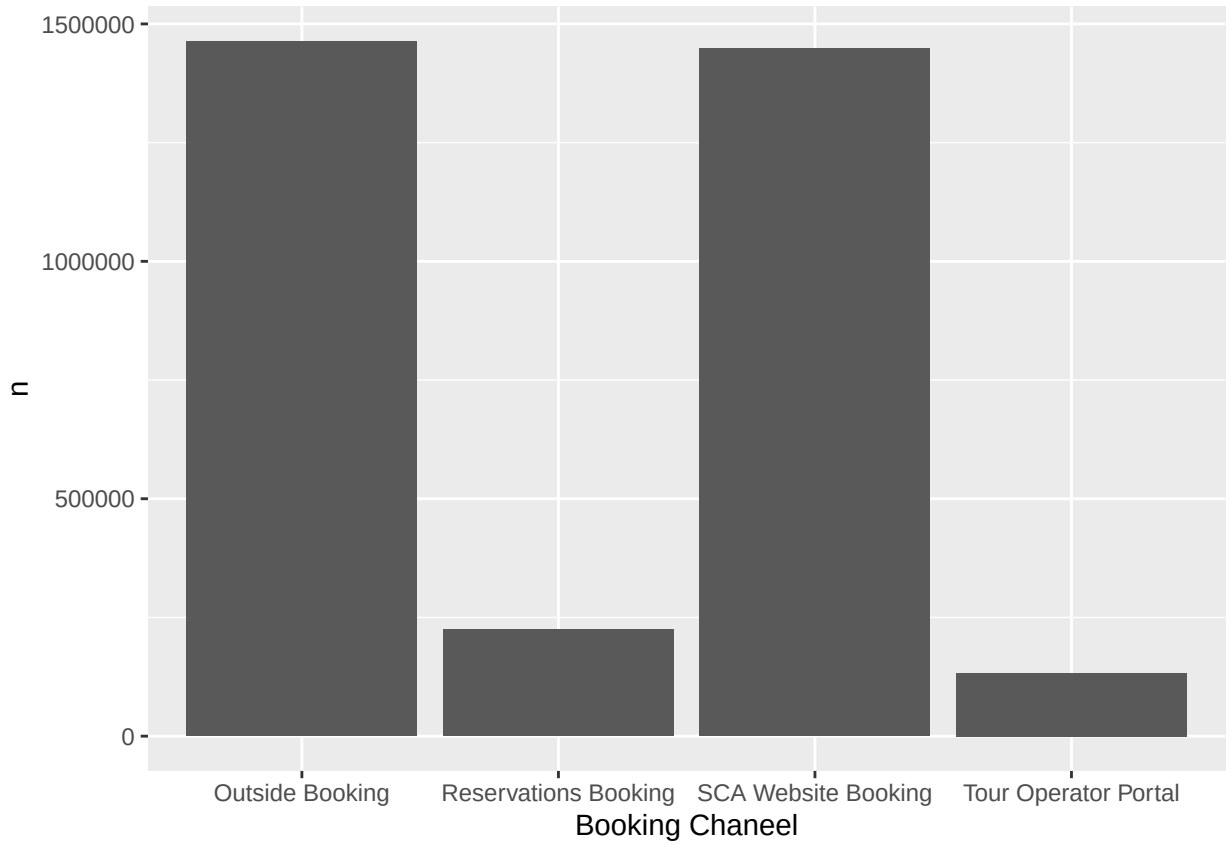
Data analysis and insight generation

In order to achieve our success and main goals, our analysis has three key objectives including : 1.Developing a deep, accurate, and robust picture of the different segments of Sun Country's customers, 2.Developing customized packages for different categories of customers 3.Ensure online booking channels meet the expectations of twenty-first century travelers

We completed these three objectives by segmenting customers using k-protptype clustering methods and then generated insights for customized packages. Finally, to accomplish the third objective, we used the clustering method of k medoids to segment the customers based on their booking behaviors.

Check the distribution of top 4 most popular booking channel.

```
df %>%
  dplyr::count(BookingChannel) %>%
  arrange(desc(n)) %>%
  head(4) %>%
  ggplot(aes(x = BookingChannel, y = n)) + geom_bar(stat = 'identity') +
  xlab('Booking Chaneel')
```



Customer Segmentation Analysis

Kprototype Clustering for segmenting customers

Selecting clsutering features and perfom random sampling of 40000 records.

We first decided on a few important features that we felt would help us better understand customer behavior. Then, we did some basic data engineering like normalization on these features to bring them to the same scale and not skew the analysis. Having these features, we select a subset of 40000 to aid the clustering analysis.

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

set.seed(1000)

df_sample1 <- df %>%
  mutate(Daysinadvance_booked_flight = as.numeric(Daysinadvance_booked_flight)) %>%
  mutate(BaseFareAmt_scaled = normalize(BaseFareAmt)) %>%
  mutate(Age_scaled = normalize(Age)) %>%
  filter(BaseFareAmt != 0) %>%
  sample_n(40000, replace = FALSE)

df_sample1_use <- df_sample1 %>%
```

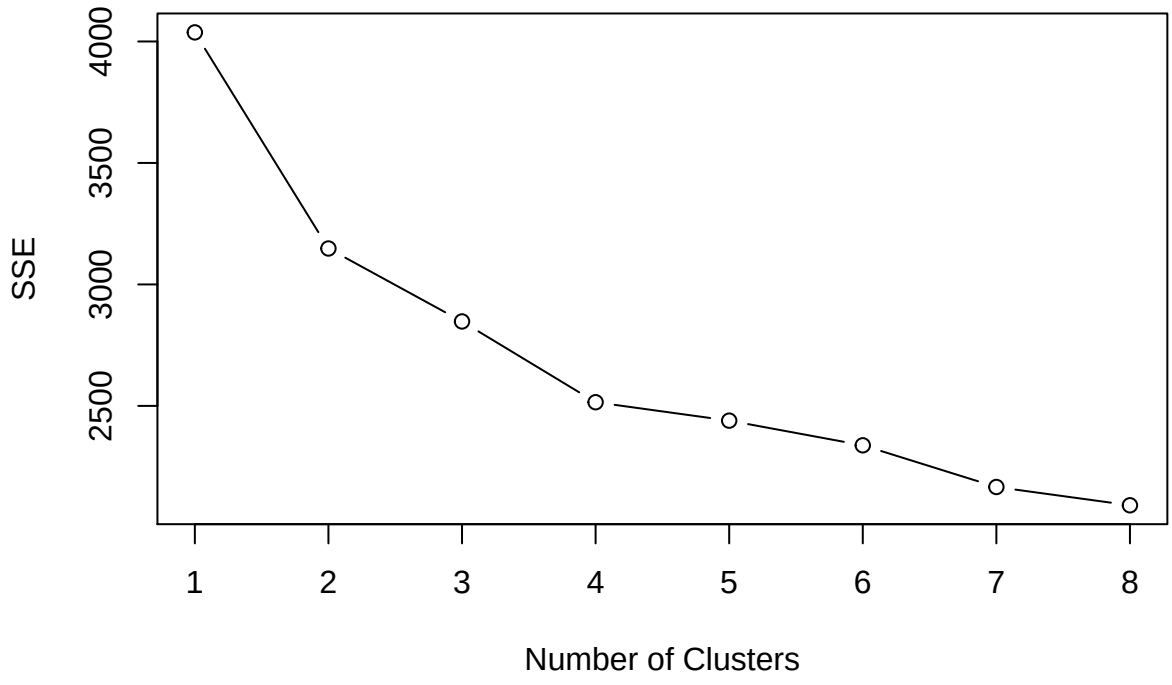
```
  select("GenderCode", "Age_scaled", "TrvldClassOfService",
         "BaseFareAmt_scaled", "Status", "FrequentFlyer", "month_flight")
```

Picking the number of clusters based on SSE, we pick 4.

From the SSE curve, we see that there is a steep drop at 4 and it gradually starts to plateau from there. So, we choose 4 clusters to be ideal to work with.

```
set.seed(9845)
SSE_curve <- c()
for (k in 1:8){
  kpro <- kproto(df_sample1_use, k)
  sse <- sum(kpro$withinss)
  SSE_curve[k] <- sse
}

plot(1:8, SSE_curve, type="b", xlab="Number of Clusters", ylab="SSE")
```



Running clustering model

We run k-prototype clustering on this sample since we have mixed data types in the features and K-prototype handles this well. Then we observe the results of how different features are distributed across the clusters and the sample size for each cluster.

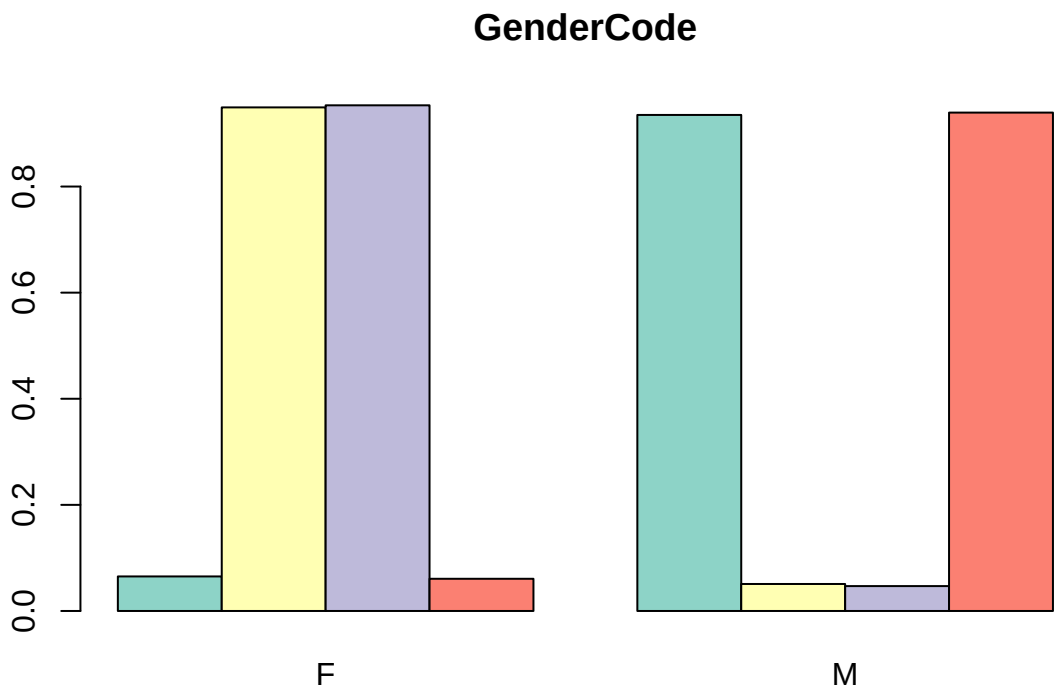
```
set.seed(1000)
sy_kproto <- df_sample1_use %>%
  clustMixType::kproto(k = 4, nstart = 30)
```

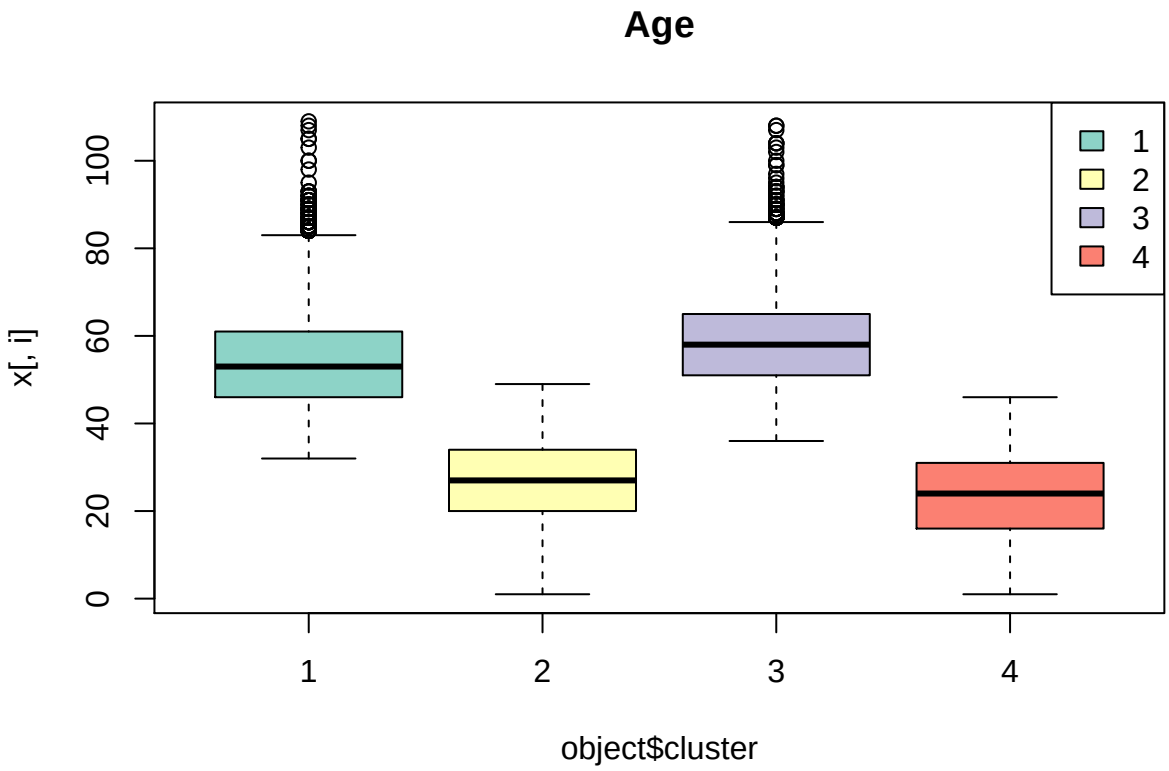
```
df_sample1_use_temp <- df_sample1 %>%
  select("GenderCode", "Age", "TrvldClassOfService",
         "BaseFareAmt", "Status", "FrequentFlyer", "month_flight")

df_sample1_use_temp <-
  df_sample1_use_temp %>%
  mutate(kproto_cluster = sy_kproto$cluster)
```

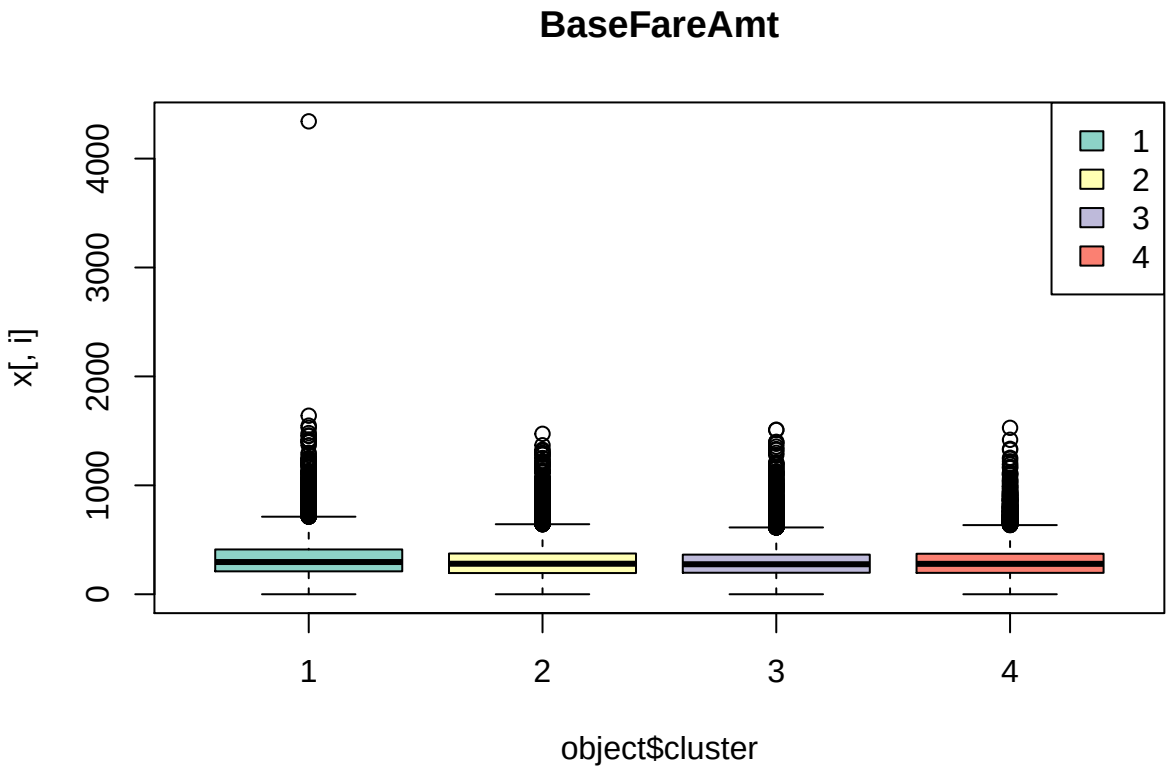
Clustering plots

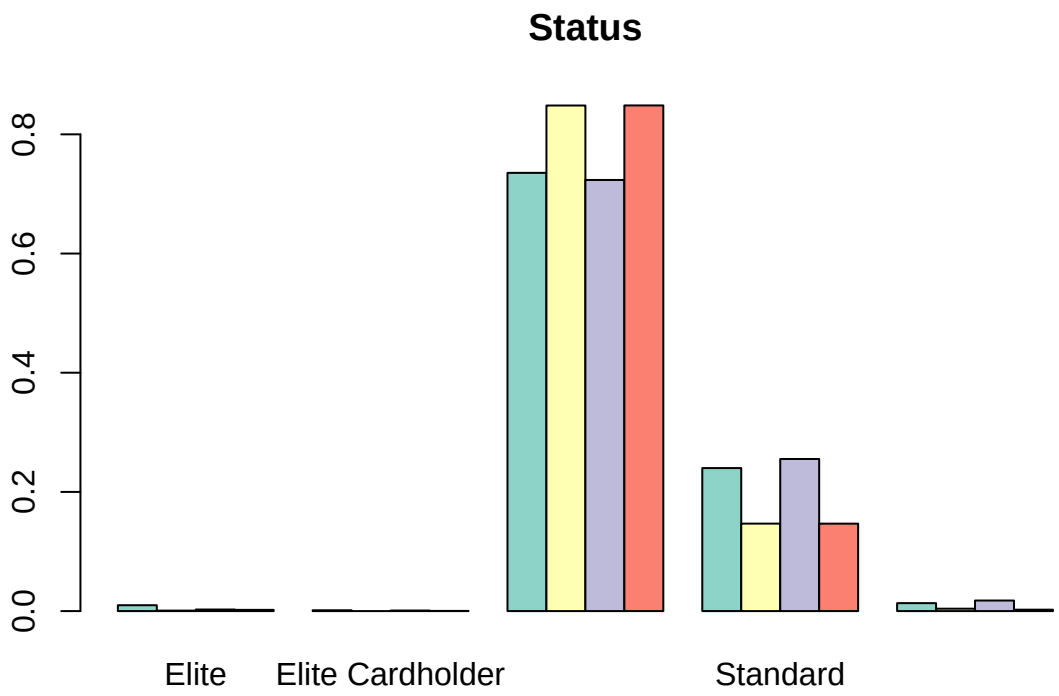
```
clprofiles(sy_kproto, df_sample1_use_temp)
```

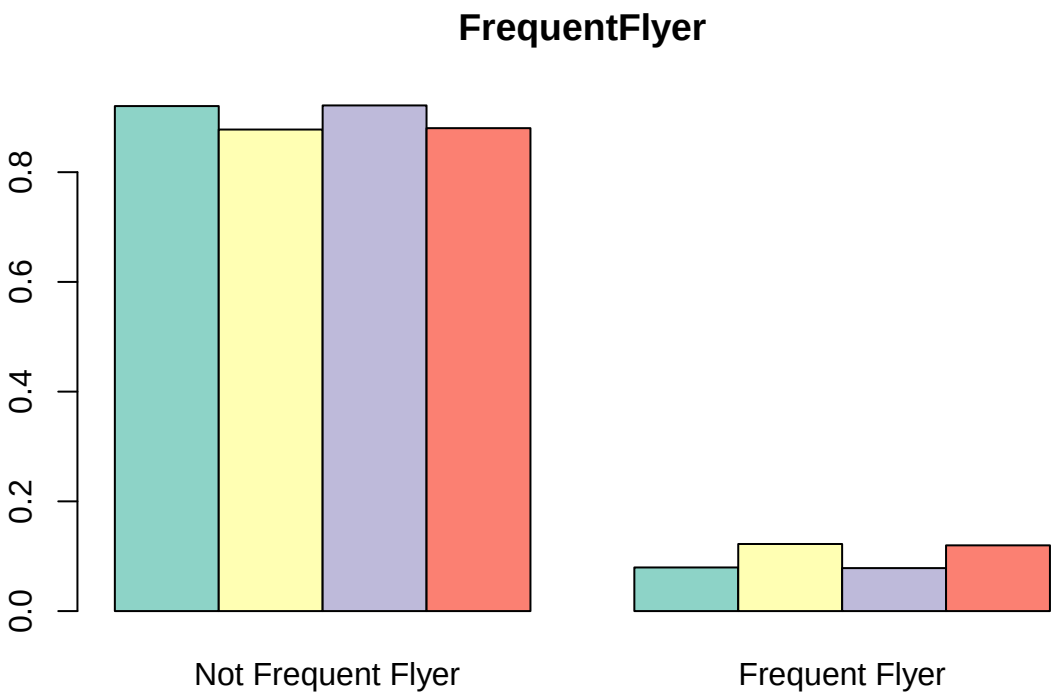


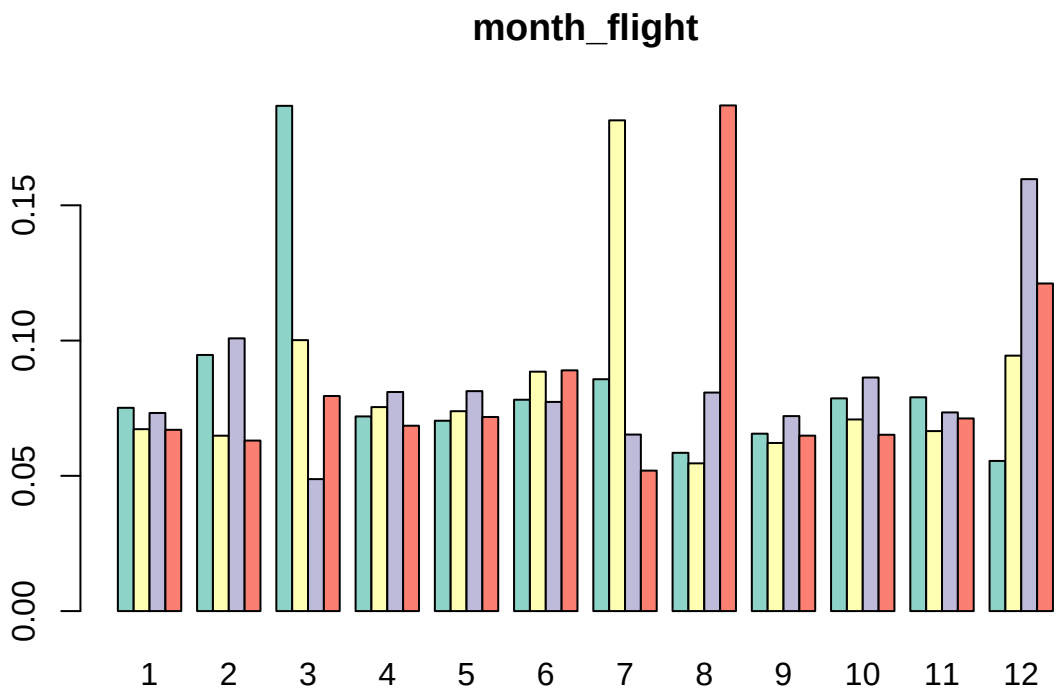

Age
x[, i]
0
20
40
60
80
100
1
2
3
4
object$cluster
1
2
3
4

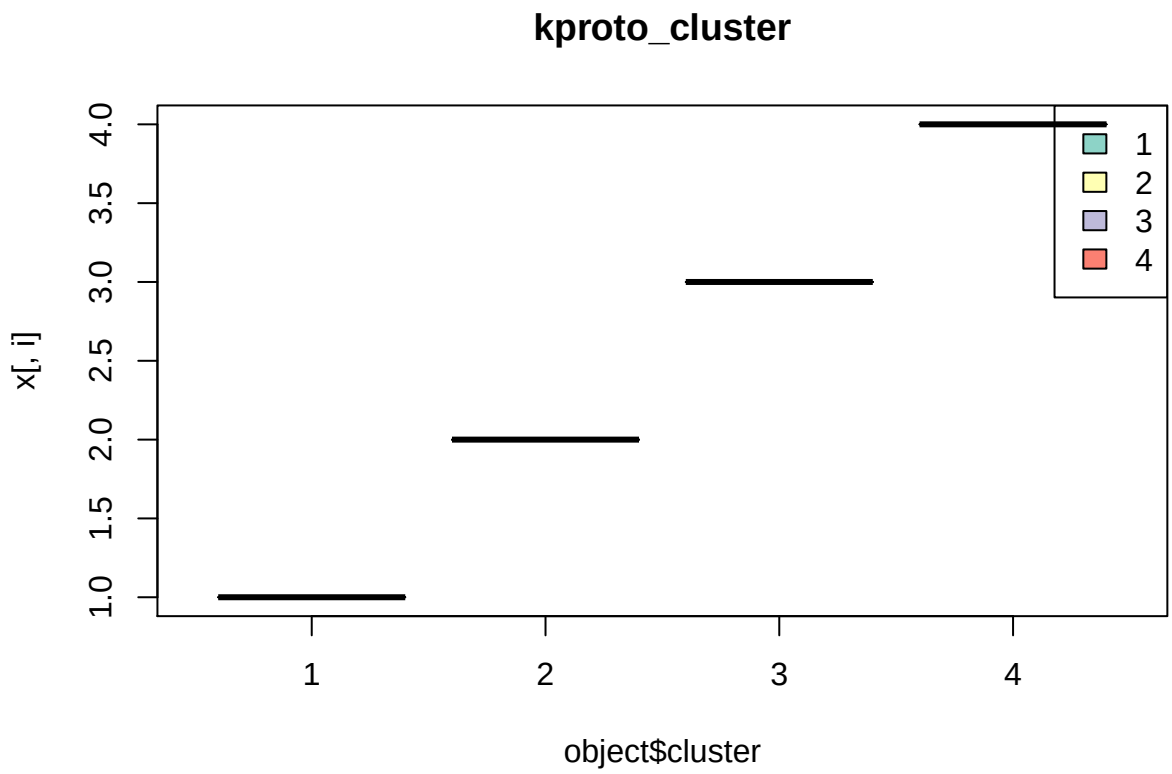

TrvldClassOfService
0.0
0.2
0.4
0.6
0.8
Coach
Discount First Class
First Class


BaseFareAmt
x[, i]
0
1000
2000
3000
4000
1
2
3
4
object$cluster
1
2
3
4


Status
0.0
0.2
0.4
0.6
0.8
Elite
Elite Cardholder
Standard


FrequentFlyer
0.0
0.2
0.4
0.6
0.8
Not Frequent Flyer
Frequent Flyer


month_flight
0.00
0.05
0.10
0.15
1
2
3
4
5
6
7
8
9
10
11
12



To better understand the result of segmentation, we can compare the clustering result with the total sample.

```
summary(sy_kproto)
```

```
## GenderCode
##
## cluster     F     M
##       1 0.065 0.935
##       2 0.949 0.051
##       3 0.953 0.047
##       4 0.061 0.939
##
## ------------------------------------------------------------
## Age_scaled
##        Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 1 0.2870370 0.4166667 0.4814815 0.4939364 0.5555556 1.0000000
## 2 0.0000000 0.1759259 0.2407407 0.2324878 0.3055556 0.4444444
## 3 0.3240741 0.4629630 0.5277778 0.5319130 0.5925926 0.9907407
## 4 0.0000000 0.1388889 0.2129630 0.2052818 0.2777778 0.4166667
##
## ------------------------------------------------------------
## TrvldClassOfService
##
## cluster Coach Discount First Class First Class
##       1 0.883                0.040       0.077
##       2 0.959                0.016       0.026
```

```
##       3 0.896                0.037       0.066
##       4 0.957                0.015       0.027
##
## ------------------------------------------------------------------
## BaseFareAmt_scaled
##           Min.    1st Qu.     Median       Mean    3rd Qu.      Max.
## 1 4.606172e-06 0.04842008 0.06812989 0.07756383 0.09469369 1.0000000
## 2 4.606172e-06 0.04485951 0.06448641 0.06944229 0.08612391 0.3392446
## 3 2.303086e-06 0.04541916 0.06341317 0.06978190 0.08376785 0.3477890
## 4 2.303086e-06 0.04520497 0.06427568 0.06994315 0.08559362 0.3519116
##
## ------------------------------------------------------------------
## Status
##
## cluster Elite Elite Cardholder Non-Member Standard Standard Carholder
##       1 0.010            0.001      0.735    0.240              0.013
##       2 0.001            0.000      0.848    0.147              0.004
##       3 0.003            0.001      0.723    0.255              0.018
##       4 0.002            0.000      0.849    0.147              0.002
##
## ------------------------------------------------------------------
## FrequentFlyer
##
## cluster  Not Frequent Flyer Frequent Flyer
##       1               0.920          0.080
##       2               0.878          0.122
##       3               0.922          0.078
##       4               0.880          0.120
##
## ------------------------------------------------------------------
## month_flight
##
## cluster     1     2     3     4     5     6     7     8     9    10    11    12
##       1 0.075 0.095 0.187 0.072 0.070 0.078 0.086 0.059 0.066 0.079 0.079 0.056
##       2 0.067 0.065 0.100 0.075 0.074 0.089 0.181 0.055 0.062 0.071 0.067 0.094
##       3 0.073 0.101 0.049 0.081 0.081 0.077 0.065 0.081 0.072 0.086 0.073 0.160
##       4 0.067 0.063 0.080 0.069 0.072 0.089 0.052 0.187 0.065 0.065 0.071 0.121
##
## ------------------------------------------------------------------
```

Descriptive statistics for the sample

```
summary(df_sample1_use)
```

```
##  GenderCode   Age_scaled              TrvldClassOfService BaseFareAmt_scaled
##  F:20886    Min.   :0.0000   Coach               :36984    Min.   :0.0000023
##  M:19114    1st Qu.:0.2315   Discount First Class: 1077    1st Qu.:0.0462782
##             Median :0.3611   First Class         : 1939    Median :0.0651313
##             Mean   :0.3631                                 Mean   :0.0716767
##             3rd Qu.:0.5000                                 3rd Qu.:0.0874090
##             Max.   :1.0000                                 Max.   :1.0000000
##
```

```
##                  Status                  FrequentFlyer    month_flight
##  Elite              :  152    Not Frequent Flyer:35979   12     : 4256
##  Elite Cardholder   :   26   Frequent Flyer     : 4021   3      : 4194
##  Non-Member         :31612                               7      : 3987
##  Standard           : 7840                               8      : 3701
##  Standard Carholder:  370                                6      : 3334
##                                                          2      : 3219
##                                                          (Other):17309
```

According to the age and gender segmentation, the four clusters are mainly composed of middle aged males, young females, middle aged females and young males. And for the entire sample data, females to males is roughly 50 to 50. Age is distributed between twenty three to fifty.

When it comes to member status and frequent flyer, the first and third cluster have higher members proportion and lower frequent flyer proportion than the entire sample dataset, while the second and fourth cluster have fewer members, but have more frequent flyers.The last column reveals that four clusters fly frequently in march, july, december and august respectively, while the flight months of the entire sample are evenly distributed among the year

Popular Destinations across clusters

For popular destinations, we take out MSP from the visualization because it is obvious MSP has the largest amount of traffic. Moreover, since Sun Country is a MSP based carrier - all their flight layovers will happen in MSP only. Hence it doesn't make sense to include MSP in the destination list as it doesnt give an idea if the people really prefer MSP or they just had a layover.

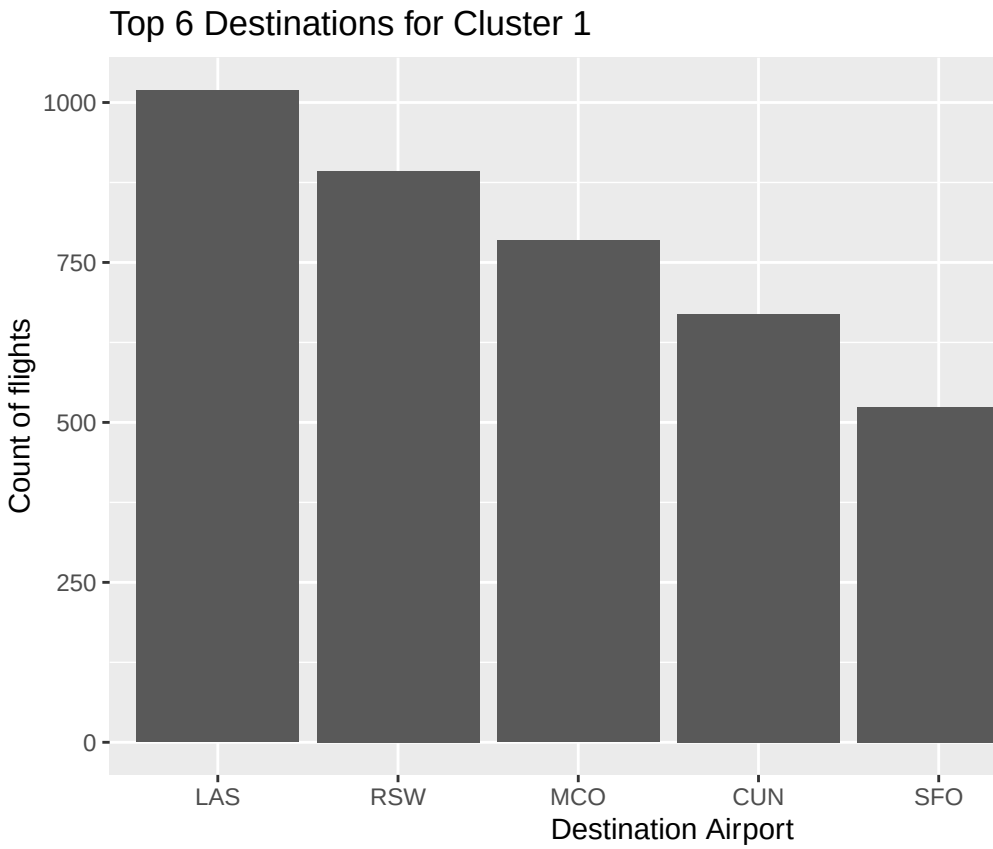
We do this to understand where people in each cluster fly to and provide targeted recommendations.

Now we try and understand the behavior of people in each cluster so that we would be able to give recommendations to target these people.

Cluster 1 Analysis

```
C1 <- df %>%
  filter(GenderCode == "M" & Status != "Non-Member") %>%
  filter(Age > 30 & Age < 50) %>%
  filter(month_flight == 3 | month_flight == 2) %>%
  group_by(ServiceEndCity) %>%
  dplyr::summarise(count = n(), .groups = 'drop') %>%
  arrange(desc(count)) %>%
  head(7)
```

```
ggplot(C1 %>%
         filter(ServiceEndCity != 'MSP'),
         aes(x = reorder(ServiceEndCity, -count), y = count)) +
         geom_bar(stat = "summary", position="dodge") +
         labs(title="Top 6 Destinations for Cluster 1",
         x=" Destination Airport", y="Count of flights")
```



Top destinations for this cluster

We notice this category of people are likely to travel during early spring (which we can observe from the clusterwise travel pattern across months in the graph mentioned in the clustering plots section above) and their top travel destinations include popular tourist locations such as Las Vegas, Florida, and Mexico.

Conclusion for Cluster 1: We could target these people by promoting customized tourist packages during spring. We also observe that this cluster has a relatively higher proportion of UFly members compared to others (from the clusterwise travel pattern across months in the graph mentioned in the clustering plots section above). We are more likely to convert these people to UFly members. So we could provide additional enrollment incentives to drive up the UFly membership.

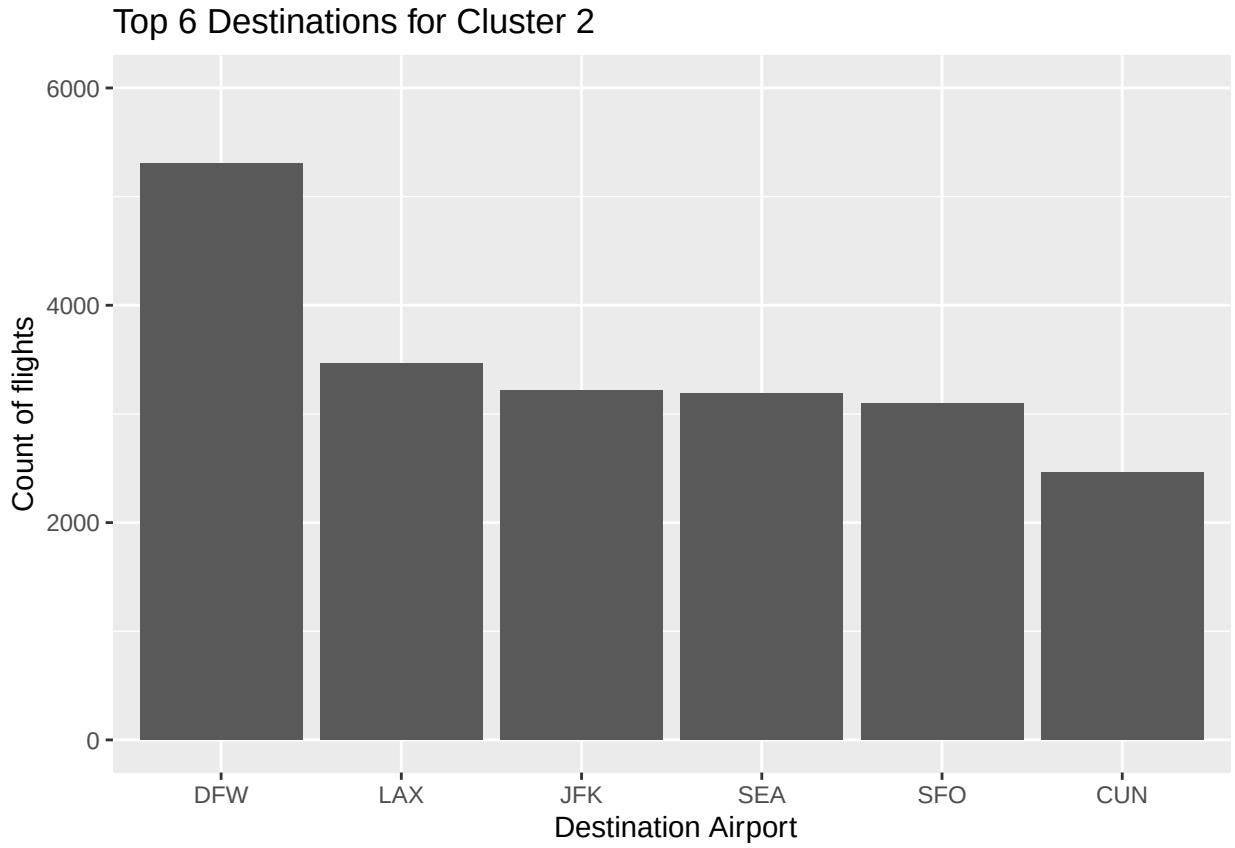
Cluster 2 Analysis

```
df$discount = factor(df$discount)

C2 <- df %>%
  filter(GenderCode == "F" & Status == "Non-Member" & Age < 30) %>%
  filter( month_flight == 7) %>%
  group_by(ServiceEndCity) %>%
  dplyr::summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(7)
```

Top destinations for this cluster

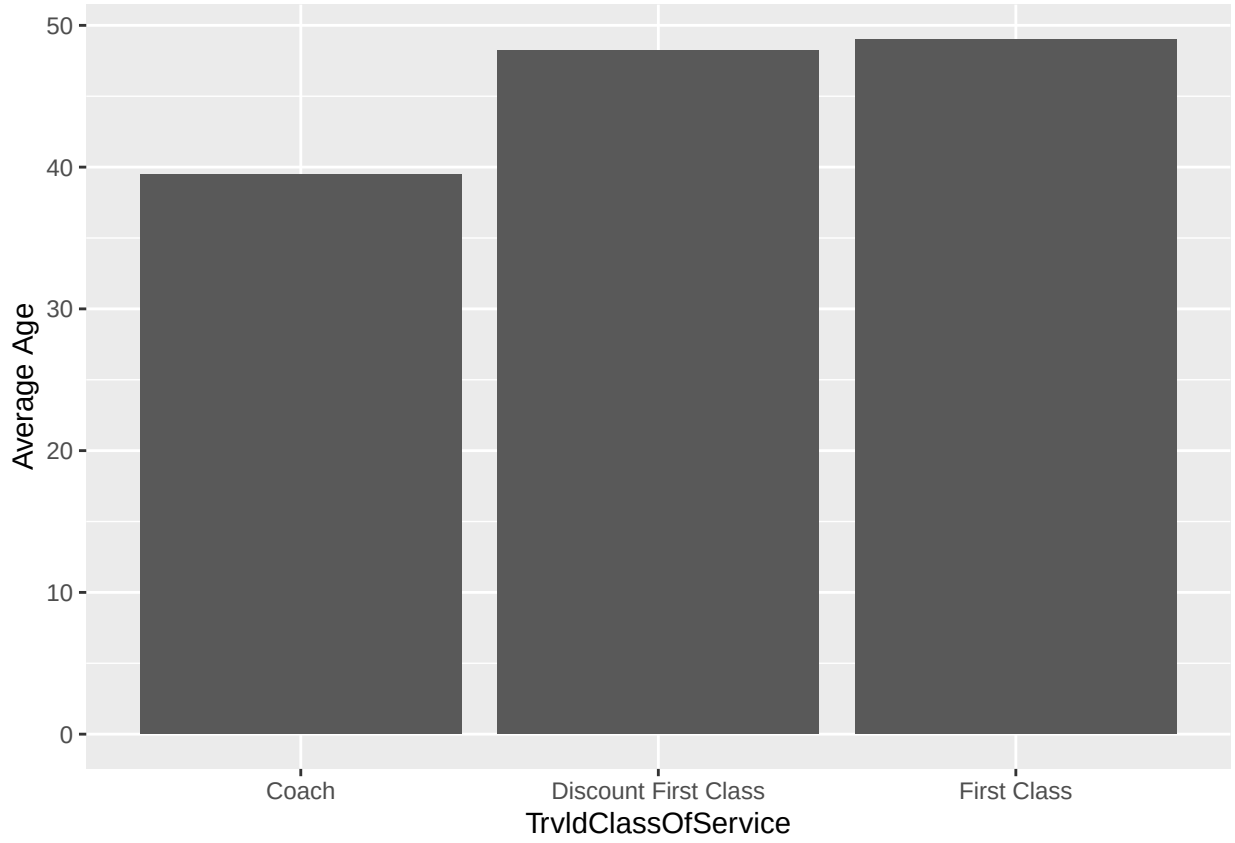
```
ggplot(C2 %>%
         filter(ServiceEndCity != 'MSP'),
         aes(x = reorder(ServiceEndCity, -count), y = count)) +
         geom_bar(stat = "summary", position="dodge") +
         labs(title="Top 6 Destinations for Cluster 2",
              x=" Destination Airport", y="Count of flights") +
         scale_x_discrete(guide = guide_axis(n.dodge=1)) +
         ylim(c(0,6000))
```



The age group of these fliers is less than 30. We also observe that fliers in this category travel the most during summer(which we can observe from the clusterwise travel pattern across months in the graph mentioned in the clustering plots section above) and travel to major cities and popular tourist hubs like Dallas, Los Angeles, New York, and Mexico

Average age across different class of travel

```
df %>%
  group_by(TrvldClassOfService) %>%
  dplyr::summarise(Avg_age = mean(Age)) %>%
  ggplot(aes(x = TrvldClassOfService, y = Avg_age)) +
  geom_bar(stat = 'identity') + ylab("Average Age")
```



T-test to test there is significant age difference between one who travel

first class and coach.

```
df_t <- df
df_t$TrvldClassOfService[which(df_t$TrvldClassOfService == "Discount First Class")] = "First Class"

t.test(Age ~ TrvldClassOfService, data = df_t)
```

```
##
##  Welch Two Sample t-test
##
## data:  Age by TrvldClassOfService
## t = -266.26, df = 310307, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.349014 -9.212381
## sample estimates:
##       mean in group Coach mean in group First Class
##                  39.46235                  48.74305
```

The results has shown that there is a significant difference between the age for Coach Class and First Class travelers.

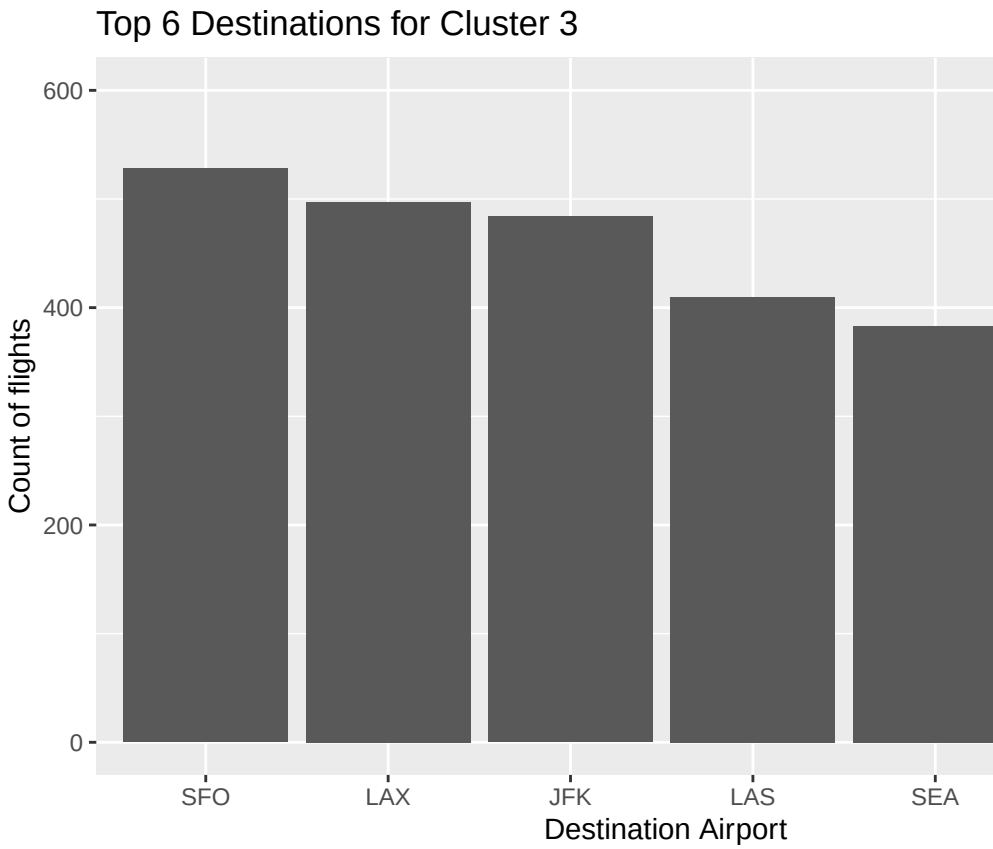
We observe that younger people are more likely to travel to coach class compared to first class.

Conclusion for Cluster 2: We could target these categories of people by providing student based offers on coach class during summer and bundle packages to popular tourist locations. This would drive more people to fly with Sun country increasing the occupancy rate of the flight and flying frequency per person, eventually contributing to increased revenue.

Cluster 3 Analysis

```
C3 <- df %>%
  filter(GenderCode == "F" & Status != "Non-Member") %>%
  filter(Age > 30 & Age < 50) %>%
  filter(month_flight == 12) %>%
  group_by(ServiceEndCity) %>%
  dplyr::summarise(count = n(), .groups = 'drop') %>%
  arrange(desc(count)) %>%
  head(7)
```

```
ggplot(C3 %>%
         filter(ServiceEndCity != 'MSP'),
         aes(x = reorder(ServiceEndCity, -count), y = count)) +
         geom_bar(stat = "summary", position="dodge") +
         labs(title="Top 6 Destinations for Cluster 3",
         x=" Destination Airport", y="Count of flights") +
         ylim(c(0,600))
```



Top destinations for this cluster

We notice that these category of people travel the most to major cities like San Francisco, Los Angeles and New york or holiday destinations like Vegas and Florida.

Impact of discount on booking class

To see the impact of discounts and plan on how to better use them, we see the impact of discounts on the class of travel being booked. We observe that, when given a discount, fliers are more likely to book first class tickets compared to booking without a discount.

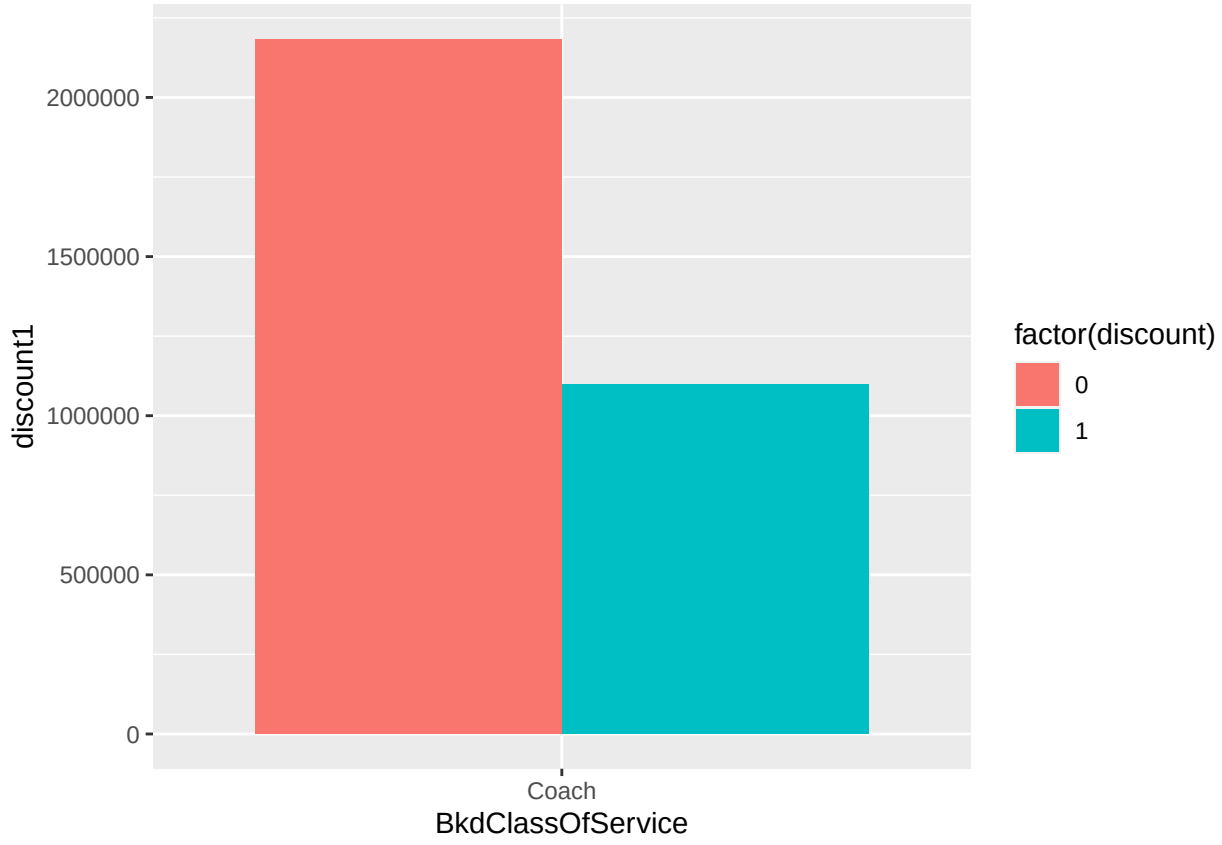
```
df_freq1 = df %>%
  group_by(discount, BkdClassOfService) %>%
  dplyr::summarise(discount1 = n())
df_freq1
```

```
## # A tibble: 6 x 3
## # Groups:   discount [2]
##   discount BkdClassOfService    discount1
##   <fct>    <fct>                    <int>
## 1 0        Coach                  2183100
## 2 0        Discount First Class       409
## 3 0        First Class               6589
## 4 1        Coach                  1098492
## 5 1        Discount First Class       399
## 6 1        First Class              85519
```

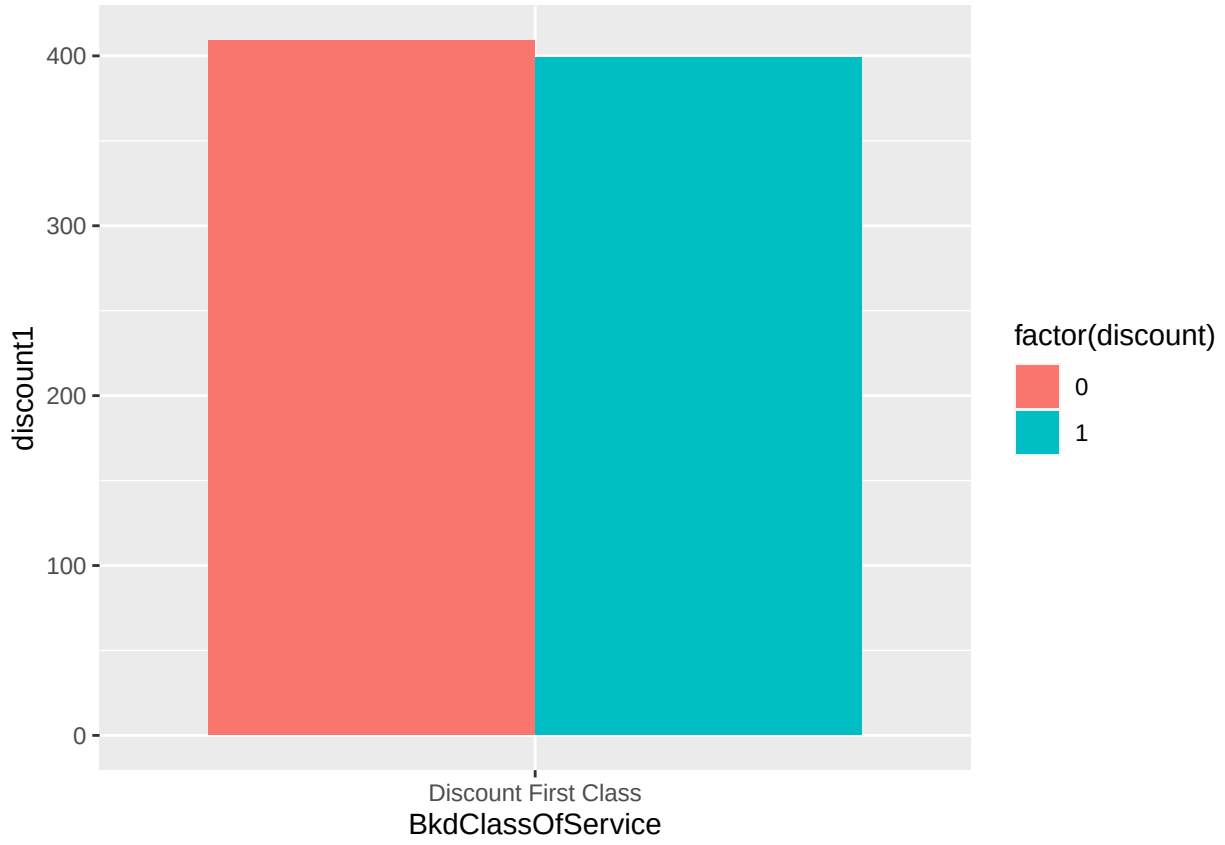
```
p1 <- ggplot(df_freq1%>%
               filter(BkdClassOfService == "Coach"),
             aes(fill=factor(discount), x=BkdClassOfService, y = discount1)) +
             geom_bar(position="dodge", stat="identity")
p2 <- ggplot(df_freq1%>%
               filter(BkdClassOfService == "Discount First Class"),
             aes(fill=factor(discount), x=BkdClassOfService, y = discount1)) +
              geom_bar(position="dodge", stat="identity")

p3 <- ggplot(df_freq1%>%
               filter(BkdClassOfService == "First Class"),
             aes(fill=factor(discount), x=BkdClassOfService, y = discount1)) +
              geom_bar(position="dodge", stat="identity")
```

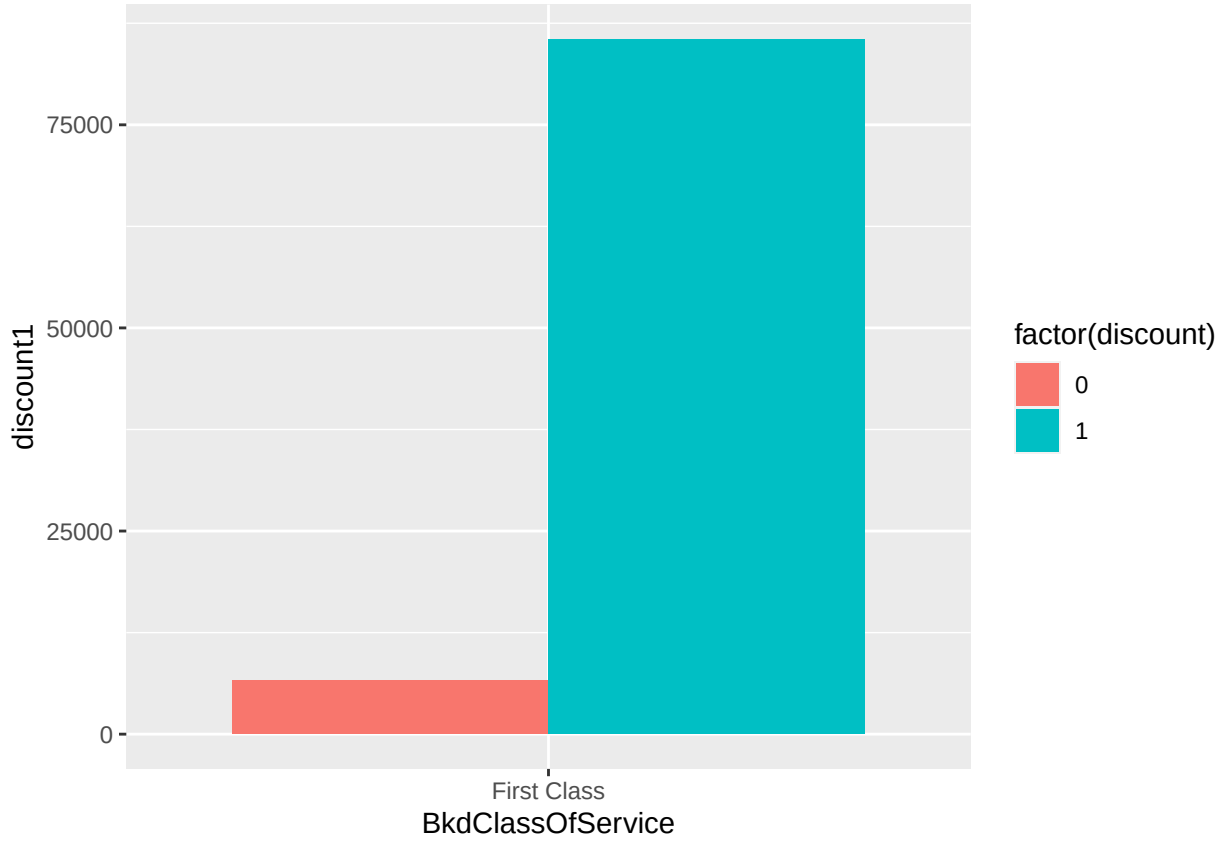
```
## We had to change plotly as it does not support pdf
##library(plotly)
##subplot(ggplotly(p1), ggplotly(p2), ggplotly(p3),nrows = 1)
par(mfrow=c(1,3))
p1
```



```
p2
```


400
300
200
100
0
discount1
Discount First Class
BkdClassOfService
factor(discount)
0
1

```
p3
```



We can observe from this general trend, we also know that providing a discount increases the likelihood of taking first class by a significant margin.

Conclusion for cluster 3: We could target this category of people by providing specialized discounts to popular destinations during winter on UFly enrollment. This could both help drive enrollment and also increase the probability of more people flying through sun country during this time.

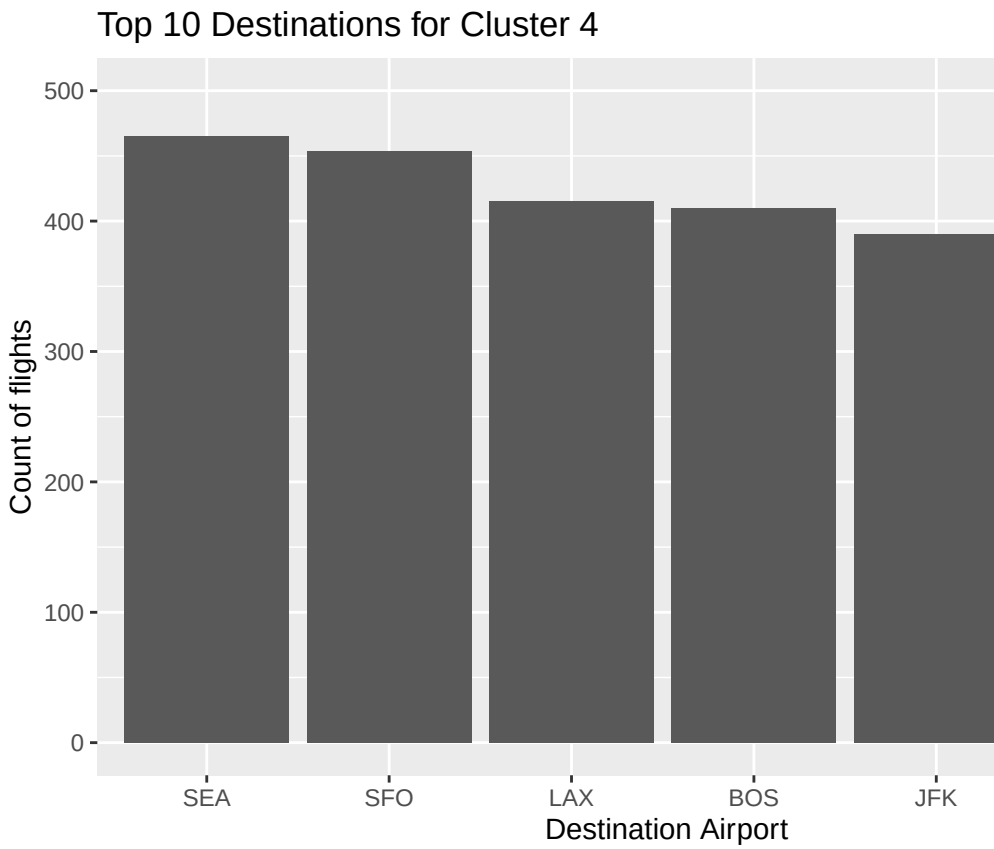
We could target these people by providing continuous discounts for first class travel in an attempt to convert more coach class customers to first class fliers thereby increasing the revenue.

Cluster 4 Analysis

```
C4 <- df %>%
  filter(GenderCode == "M" & Status != "Non-Member") %>%
  filter(Age < 30) %>%
  filter(month_flight == 8) %>%
  group_by(ServiceEndCity) %>%
  dplyr::summarise(count = n(), .groups = 'drop') %>%
  arrange(desc(count)) %>%
  head(7)
```

```
ggplot(C4 %>%
         filter(ServiceEndCity != 'MSP'),
```

```
        aes(x = reorder(ServiceEndCity, -count), y = count)) +
        geom_bar(stat = "summary", position="dodge") +
        labs(title="Top 10 Destinations for Cluster 4",
        x=" Destination Airport", y="Count of flights") +
        ylim(c(0,500))
```



Top destinations for this cluster

We observed that these category of people travel the most during early fall especially to major cities like San Francisco, Los Angeles, Seattle, Boston and New York.

First class break down to study the demographic who often purchase first class

We group fliers into different age groups and observe which class of travel each age group prefers.

Fist class break down to study the demographic who often purchase first class

```
df_ages = df %>% mutate(Age_disc = ifelse(Age <= 30, "Young",
                            ifelse(Age > 30 & Age <=50,"Middle Age", "Old")))

ggplot(df_ages %>%
        filter(TrvldClassOfService != 'Coach') %>%
        group_by(Age_disc, TrvldClassOfService, GenderCode) %>%
        dplyr::summarise(count = n()), aes(x=TrvldClassOfService, y=count,
                                    groups=Age_disc, fill=Age_disc)) +
```

```
        geom_bar(stat='identity', position='dodge')  +
        labs(title="First Class Breakdown by Gender and Age",
             x="Class", y="Number of Flights", fill='Age Groups') +
        facet_wrap(GenderCode~.)
```

We observed that younger people tend to prefer coach class as compared to older people.

Conclusion for Cluster 4: We could target this category by providing discounts to their top destinations during fall for coach class. Additionally, we can also provide them with bonus miles to increase their flying frequency. Moreover, people in this group tend to have a lot of contacts with similar demographics and by providing referral bonus Sun Country will be able to increase their customer acquisition.

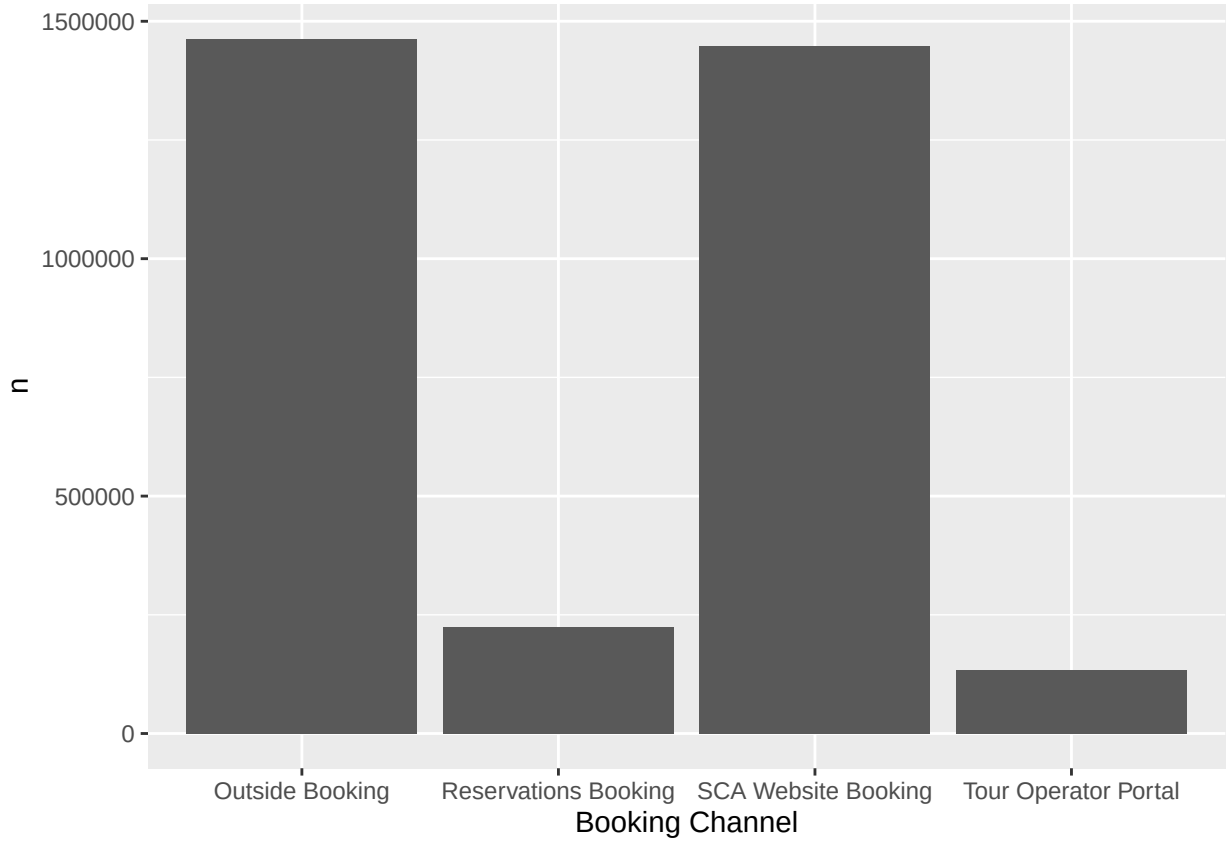
Booking Channel Analysis

Check the distribution of the top 4 most popular booking channels.

We understand the channel the fliers use to book their tickets. We do this to understand what proportion of them are already using online booking channels and how we could further drive them. We also explore the characteristics of fliers who are making online booking and make recommendations to target them and also drive website usage.

```
df %>%
  dplyr::count(BookingChannel) %>%
  arrange(desc(n)) %>%
  head(4) %>%
```

```
ggplot(aes(x = BookingChannel, y = n)) + geom_bar(stat = 'identity') +
xlab('Booking Channel')
```



To better understand the customer behavior related to the website booking, we select the customer booking channel ="SCA Website Booking". After preprocessing the data, we used the K-medoids clustering to segment the customer into 5 groups based on the elbow curve.

```
df=df%>%unite("Unique_ID", EncryptedName:birthdateid, remove = FALSE)
df$GenderCode=as.factor(df$GenderCode)
df$BookingChannel=as.factor(df$BookingChannel)
df$Status=as.factor(df$Status)
df$TrvldClassOfService=as.factor(df$TrvldClassOfService)
df$Stopover=as.factor(df$Stopover)
df$Daysinadvance_booked_flight=as.numeric(df$Daysinadvance_booked_flight)
df$memebership_duration_before_flight=as.numeric(df$memebership_duration_before_flight)
options(scipen = 100)
```

Set seed and choose the variable to use, unique id is to identify the customer and TrvldClassOfService,Daysinadvance_booked_flight are the booking related behaviors

```
set.seed(5)
df_web<-df%>%filter(BookingChannel=="SCA Website Booking")
df_web<-df_web%>%sample_n(15000,replace=FALSE)
df_cluster=df_web%>%select(Unique_ID,Age,TrvldClassOfService,Daysinadvance_booked_flight)
glimpse(df_cluster)
```

```
## Rows: 15,000
## Columns: 4
## $ Unique_ID                   <chr> "50414C4D455244696420493F7C206765742074686~
## $ Age                         <int> 55, 18, 14, 86, 57, 50, 38, 73, 3, 42, 36,~
## $ TrvldClassOfService         <fct> Coach, Coach, Coach, Coach, Coach, Coach, ~
## $ Daysinadvance_booked_flight <dbl> 151, 7, 70, 23, 22, 45, 43, 15, 39, 124, 2~
```

First we calculate the gower distance among customers to handle mixed types of data.

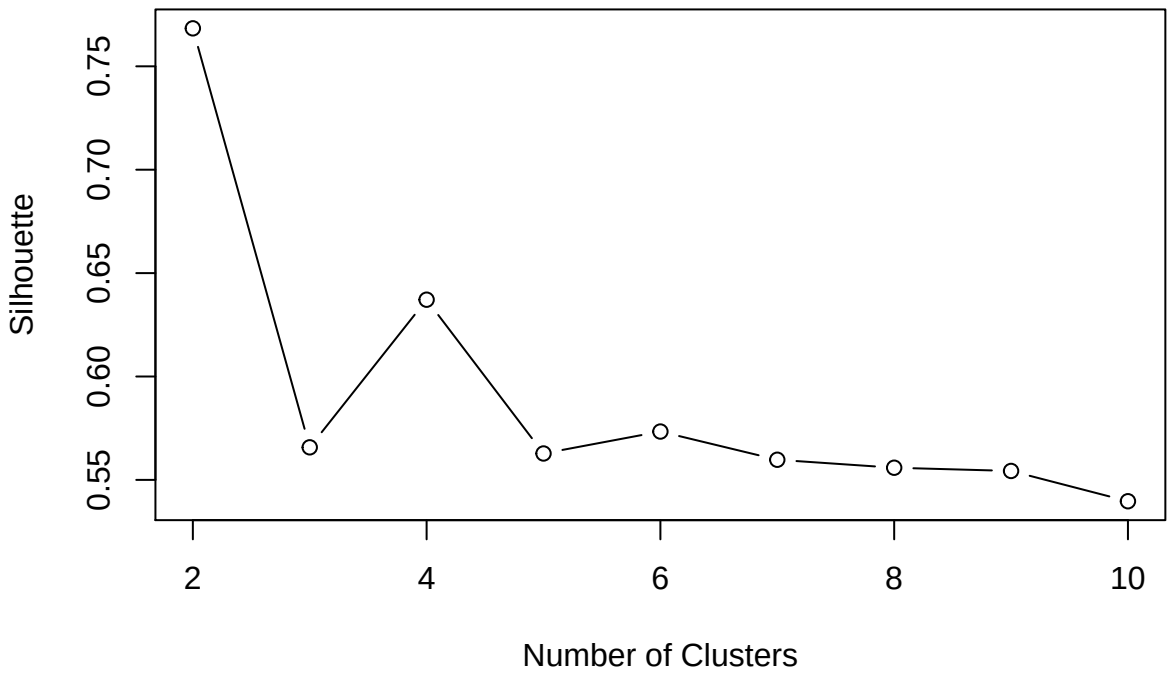
```
library(cluster) # for similarity and pam
library(ggplot2) # for visualization
## calculating gower distance
gower_dist <- daisy(df_cluster[, -1],
                    metric = "gower",
                    type = list(logratio = 3))
summary(gower_dist)
```

```
## 112492500 dissimilarities, summarized :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.05556 0.12626 0.20498 0.24242 1.00000
## Metric :  mixed ;  Types = I, N, I
## Number of objects : 15000
```

```
gower_mat <- as.matrix(gower_dist)
```

We draw a silhouette curve to determine the k value, and it appears that the cutoff is 5 based on the elbow curve.

```
# to determine proper k value
sil_curve <- c()
for (k in 2:10) {
  pam_fit <- pam(gower_dist, diss = TRUE,k = k)
  #PAM interally computes the silhouette measure
  sil_curve[k] <- pam_fit$silinfo$avg.width
}
sil_curve = sil_curve[2:10]
plot(2:10, sil_curve, type="b", xlab="Number of Clusters", ylab="Silhouette")
```




```
num_pam_clusters = which.max(sil_curve)+1
```

K-mediods clustering

K-medoids clustering result summary, we can generate the cluster size and observe each cluster situation

```
pam_fit <- pam(gower_dist, diss = TRUE, k =5)

pam_results <- df_cluster %>%
  mutate(cluster = pam_fit$clustering) %>%
  group_by(cluster) %>%
  do(the_summary = summary(.))

### clustering result summary
pam_results$the_summary
```

```
## [[1]]
##   Unique_ID              Age                    TrvldClassOfService
##  Length:4320        Min.   : 54.00   Coach               :4320
##  Class :character   1st Qu.: 58.00   Discount First Class:   0
##  Mode  :character   Median : 63.00   First Class         :   0
##                     Mean   : 64.03
##                     3rd Qu.: 68.00
##                     Max.   :100.00
```

```
##  Daysinadvance_booked_flight    cluster
##  Min.   :  0.00              Min.   :1
##  1st Qu.: 27.00              1st Qu.:1
##  Median : 51.00              Median :1
##  Mean   : 66.71              Mean   :1
##  3rd Qu.: 90.00              3rd Qu.:1
##  Max.   :710.00              Max.   :1
##
## [[2]]
##   Unique_ID              Age                  TrvldClassOfService
##  Length:4949        Min.   : 1.00   Coach               :4949
##  Class :character   1st Qu.:13.00   Discount First Class:   0
##  Mode  :character   Median :22.00   First Class         :   0
##                     Mean   :19.87
##                     3rd Qu.:28.00
##                     Max.   :32.00
##  Daysinadvance_booked_flight    cluster
##  Min.   :  0.00              Min.   :2
##  1st Qu.: 26.00              1st Qu.:2
##  Median : 50.00              Median :2
##  Mean   : 62.76              Mean   :2
##  3rd Qu.: 84.00              3rd Qu.:2
##  Max.   :604.00              Max.   :2
##
## [[3]]
##   Unique_ID              Age                  TrvldClassOfService
##  Length:4263        Min.   :33.00   Coach               :4263
##  Class :character   1st Qu.:38.00   Discount First Class:   0
##  Mode  :character   Median :44.00   First Class         :   0
##                     Mean   :43.45
##                     3rd Qu.:49.00
##                     Max.   :53.00
##  Daysinadvance_booked_flight    cluster
##  Min.   :  0.00              Min.   :3
##  1st Qu.: 24.00              1st Qu.:3
##  Median : 48.00              Median :3
##  Mean   : 63.21              Mean   :3
##  3rd Qu.: 85.00              3rd Qu.:3
##  Max.   :516.00              Max.   :3
##
## [[4]]
##   Unique_ID              Age                 TrvldClassOfService
##  Length:975         Min.   : 1.0   Coach               :  0
##  Class :character   1st Qu.:39.0   Discount First Class:  0
##  Mode  :character   Median :53.0   First Class         :975
##                     Mean   :50.3
##                     3rd Qu.:62.5
##                     Max.   :95.0
##  Daysinadvance_booked_flight    cluster
##  Min.   :  0.00              Min.   :4
##  1st Qu.: 17.00              1st Qu.:4
##  Median : 39.00              Median :4
##  Mean   : 57.67              Mean   :4
##  3rd Qu.: 75.00              3rd Qu.:4
```

```
##  Max.   :649.00              Max.   :4
##
## [[5]]
##   Unique_ID              Age                  TrvldClassOfService
##  Length:493         Min.   : 1.00   Coach               :  0
##  Class :character   1st Qu.:39.00   Discount First Class:493
##  Mode  :character   Median :50.00   First Class         :  0
##                     Mean   :49.08
##                     3rd Qu.:60.00
##                     Max.   :87.00
##  Daysinadvance_booked_flight    cluster
##  Min.   :  0.00              Min.   :5
##  1st Qu.: 23.00              1st Qu.:5
##  Median : 48.00              Median :5
##  Mean   : 63.12              Mean   :5
##  3rd Qu.: 84.00              3rd Qu.:5
##  Max.   :421.00              Max.   :5
```

According to our exploration, we found that over two-thirds of the bookings through websites are done by young or middle-aged fliers.

Introducing an app version of the website and providing additional flying incentives for bookings through the website could drive website usage. To further understand this demographic, we implemented clustering on a sample of the data and it gave us the following 5 clusters. We came up with cluster-level recommendations to improve the overall website experiences and revenues.

#the size of each cluster

```
dfM=df_cluster[pam_fit$medoids, ]
dfM
```

```
##                                                                                                  Unique_ID
## 14997 43414C4C494E414E44696420493F7C20676574207468697320726967687444 4F4E4E41204D41524945_F_36120
## 14953                          5452414E44696420493F7C20676574207468697320726967687454414D_F_51113
## 14986                 57454C434844696420493F7C20676574207468697320726967687444415 74E204A_F_43603
## 580                 42494E4445524469642 0493F7C20676574207468697320726967687 44D41524B2057_M_40233
## 14460         4841525249534F4E44696420493F7C20676574207468697320726967687 44D49434841454C_M_41419
##       Age  TrvldClassOfService Daysinadvance_booked_flight
## 14997  63                Coach                          37
## 14953  22                Coach                          25
## 14986  44                Coach                          27
## 580    53          First Class                           9
## 14460  50 Discount First Class                         190
```

To target fliers over age 60, we can give senior citizens incentives to increase their flying frequency. We can target young and middle aged fliers through coach based discounts a month in advance to increase their flying frequency.

For people who book first class around 10 days in advance and attend business trips, we can apply surge-pricing to increase fare as they are going to fly regardless of the rate.

For people who book long in advance, Sun Country can offer early discounts for first class to convert coach flyers to first-class flyers to increase revenue.

Conclusion Recommendations

Take advantage of customer segmentations by offering targeted ads or deals

1.Promote tourist packages during spring for this category; since cluster 1 has a higher likelihood for UFly enrollment, target this group by providing additional enrollment incentives

2.Fliers in the age range of 17-30 are more likely to be students; Target this group using student discounts for coach class during summer and bundle packages to popular tourist locations

3.Offer specialized discounts for travel to popular holiday destinations during winter on UFly enrollment; Provide additional discounts for first class travel as people in this category have more chance of taking a first class trip when given a discount, eventually leading to increased revenue

4.Provide discounts for top destinations during the fall for coach class; Provide bonus miles to increased flying frequency and referral bonus to increase customer acquisition

Offer advertisements to specific age groups at different times

before a flight through website

1.Working/student population in the age group of 18-45 choose coach class and book around a month in advance - Provide early booking discounts and student discounts for these category of people

2.People in the age range 50-60 booking only 10 days in advance are more likely to travel for business - These would be mandatory trips and could still generate revenue without discounts

3.General middle aged people book tickets in advance and could be targeted by early booking offers and vacation packages - we could convert them to first class by providing discount