Final Project Part 2

Nathan Miller, Eli Schultz, Shannon Takahata

11/26/2021

East 6

Introduction

Airport executives and state and local officials are curious about customer satisfaction at San Francisco International Airport.Identifying current strengths of the airport and areas for improvement are critical to increasing traffic and revenue.

A survey was administered to 3,234 participants in September 2010 and contains 100 different variables. The results contain two identifiers and mostly nominal and ordinal categorical variables. Nominal responses are for questions on topics such as airline, services used, comments grouped into areas of improvement or positive feedback, destination, and more. Ordinal responses cover levels of satisfaction on various areas or services, ease of use, and size of destination airport. A few binned numeric responses are also included, such as income, age, destination airport size, and time of flight. The only continuous variable is the weight assigned to each response by percent of passengers by terminal.

Part A

Data Wrangling

Read in the data.

Re-name variable of interest, remove variables that are not of interest, and re-code text variable to actual coded responses.

```
oldnames = c("Q6A",
             "Q6B",
             "Q6C",
             "Q6D",
             "Q6E",
             "Q6F",
             "Q6G",
             "Q6H",
             "Q6I",
             "Q6J",
             "Q6K",
             "Q6L",
             "Q6M",
             "Q6N",
             "Q7COM1",
             "Q7COM2",
             "Q7COM3",
             "Q8A",
             "Q8B",
             "Q8C",
             "Q8D",
             "Q8E",
             "Q8F",
             "Q9",
             "Q10",
             "Q11",
             "Q11A1",
             "Q11A2",
             "Q11A3",
             "Q17",
             "Q18",
             "Q19",
             "DESTMARK",
             "DESTGEO",
             "Q7_text_All")
newnames = c("Artwork_and_exihibits_0_6",
             "Restaurants_0_6",
             "Retail_shops_0_6",
             "Signs_and_directions_inside_SFO_0_6",
             "Escalators_elevators_moving_walkways_0_6",
             "Information_on_screen_monitors_0_6",
             "Information_booths_lower_level_0_6",
             "Information_booths_upper_level_0_6",
             "Signs_and_directsion_on_roadways_0_6",
             "Airport_parking_facilities_0_6",
             "AirTrain_0_6",
             "Long_term_parking_lot_shuttle_0_6",
             "Airport_rental_car_center_0_6",
             "CFO_Airport_as_a_whole_0_6",
             "Com_For_Improve1",
             "Com_For_Improve2",
             "Com_For_Improve3",
             "Cleanliness_Boarding_area_0_6",
             "Cleanliness_Parking_Garage_0_6",
             "Cleanliness_Airtrain_0_6",
             "Cleanliness_Airport_rental_car_center_0_6",
             "Cleanliness_Airport_Restaurants_0_6",
             "Cleanliness_Restrooms_0_6",
             "Safety_of_SFO_1_5",
             "Used_website",
             "Usefulness_of_website",
             "Com_Useful_feature_of_website_1",
             "Com_Useful_feature_of_website_2",
             "Com_Useful_feature_of_website_3",
             "Age",
             "Gender",
             "Income_code",
             "Destination_Market",
             "Destination_Geo_Loc",
             "Comments_For_Improvement_Txt")

## Rename and recode comments for improvements and web features
SelectedData <- SFOdata %>%
  rename_at(vars(all_of(oldnames)), ~ newnames) %>%
  dplyr::select(all_of(newnames), RESPNUM) %>%
  left_join(CommentForImprovement, by = c("Com_For_Improve1" = "Key")) %>%
  mutate(Com_For_Improve1 = CommentForImprovement) %>%
  dplyr::select(-CommentForImprovement) %>%
  left_join(CommentForImprovement, by = c("Com_For_Improve2" = "Key")) %>%
  mutate(Com_For_Improve2 = CommentForImprovement) %>%
  dplyr::select(-CommentForImprovement) %>%
  left_join(CommentForImprovement, by = c("Com_For_Improve3" = "Key")) %>%
  mutate(Com_For_Improve3 = CommentForImprovement) %>%
  dplyr::select(-CommentForImprovement) %>%
  left_join(UsefulWebFeatures, by = c("Com_Useful_feature_of_website_1" = "key")) %>%
  mutate(Com_Useful_feature_of_website_1 = WebFeature) %>%
  dplyr::select(-WebFeature) %>%
  left_join(UsefulWebFeatures, by = c("Com_Useful_feature_of_website_2" = "key")) %>%
  mutate(Com_Useful_feature_of_website_2 = WebFeature) %>%
  dplyr::select(-WebFeature) %>%
  left_join(UsefulWebFeatures, by = c("Com_Useful_feature_of_website_3" = "key")) %>%
  mutate(Com_Useful_feature_of_website_3 = WebFeature) %>%
  dplyr::select(-WebFeature) %>%
  mutate(Gender = ifelse(Gender == 0, NA, Gender - 1))
```

Encode missing data, responses that were originally at extreme ends of an ordinal scale were actually non applicable, or non-response, so re-coding those variable with explicit NA makes the resulting data more manageable.

```
missing_encoded <- SelectedData %>%
  dplyr::select(-starts_with("Com_"),-Usefulness_of_website) %>%
#  mutate_at( vars(ends_with("0_6")),~as.factor(.)) %>%
  mutate_at( vars(ends_with("0_6")),~car::recode(.,"6= NA;0= NA")) %>%
  mutate_at( vars(ends_with("Age")),~car::recode(.,"8= NA;0= NA")) #recoded 8 to be NA
```

Question 1

Customers were asked to rate their opinion of the “SFO Airport as a whole” on a scale from 1 (“unacceptable”) to 5 (“outstanding”). The executives want to know if there are patterns across the satisfied or dissatisfied customers based on demographic characteristics, such as sex, age group, and income level.

We can see how the individual demographics (age, income, gender) relate to the overall satisfaction through plotting the individual demographics. To determine if there are patterns across the level of satisfaction based on demographic characteristics, we can use clustering for our analysis so that we can see how the different customers are grouped.

Clustering allows us to combine the different demographics and then group similar objects together. Cluster models assign observations based on the distance of a central point (such as the mean). Since we can calculate the mean of our variable (the responses are non-categorical 1-5 rating), we can use the mean as the representative average of our cluster. We can use this unsupervised machine learning method to classify the observations to find the different categories of the SFO customers.

We will first try to understand the composition of airport ratings by each demographic dimension.

```
#select the demographics of interest
q1 <- missing_encoded %>%
  dplyr::select(Age,
         Income_code,
         Gender,
         CFO_Airport_as_a_whole_0_6)
q1_factors <- q1 %>%
  mutate(Age = as.factor(Age),
         Income_code = as.factor(Income_code),
         Gender = as.factor(Gender))
```

1. Age:

From the histogram, the age group 25-34 filled out the most survey. Ratings of 4 were the most prevalent across all age groups.

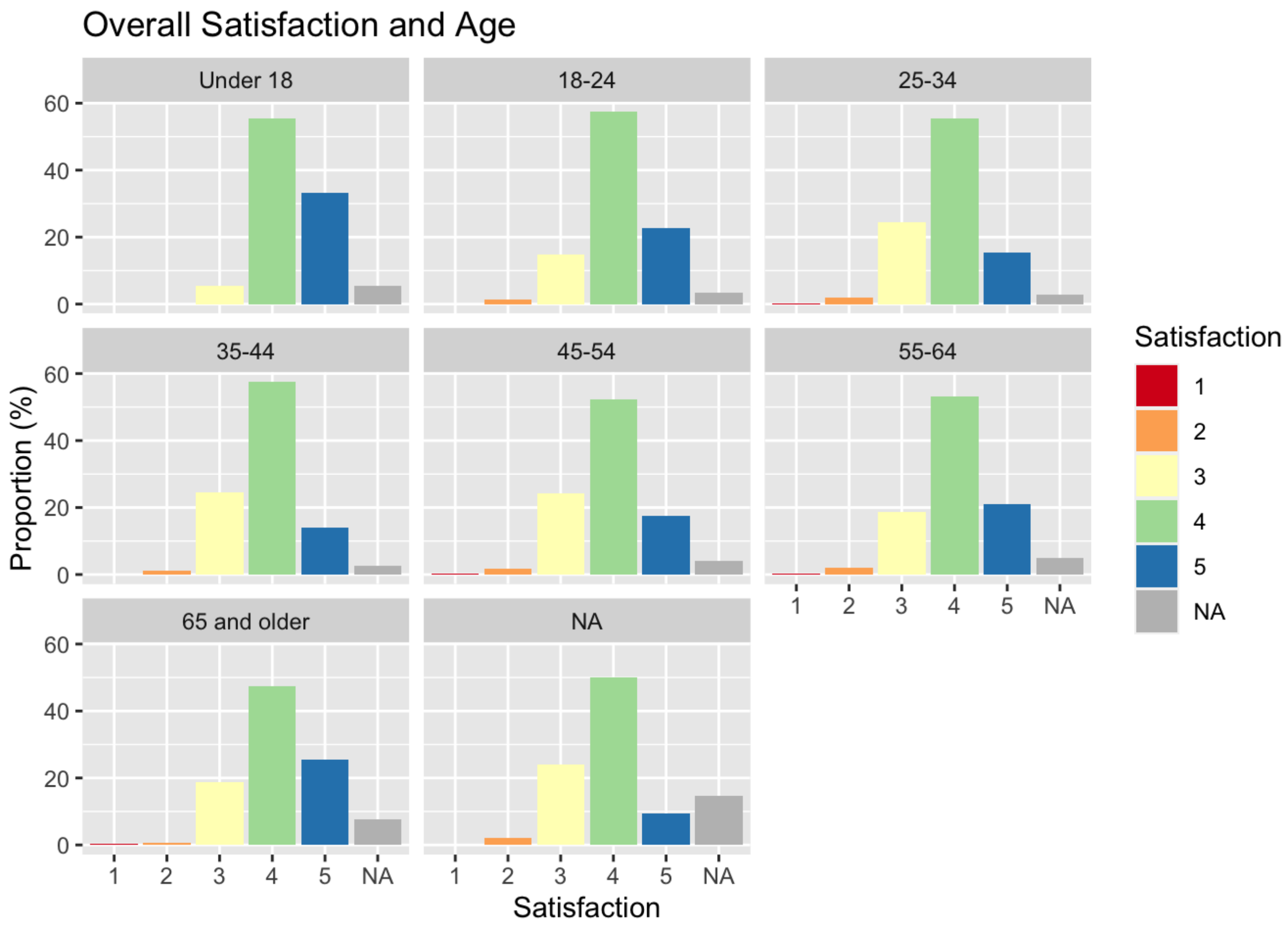
Created a subset of data with the overall customer satisfaction scores and the demographic variables.

```
# Give names to Age
levels(q1_factors$Age) <- c("Under 18","18-24","25-34","35-44","45-54","55-64","65 and o
lder")

q1_prop <- q1_factors %>%
  group_by(Age) %>%
  summarise(n_age = n()) %>%
  left_join(q1_factors %>%
              group_by(Age, CFO_Airport_as_a_whole_0_6) %>%
              summarise(n_sd_inside_sat = n())) %>%
  mutate(age_prop = (n_sd_inside_sat / n_age) * 100)
```

```
## `summarise()` has grouped output by 'Age'. You can override using the `.groups` argum
ent.
```

```
## Joining, by = "Age"
```

```
ggplot(data = q1_prop, aes(x= as.factor(CFO_Airport_as_a_whole_0_6), y = age_prop, fill
= as.factor(CFO_Airport_as_a_whole_0_6))) +
  geom_col(position = 'dodge') +
  facet_wrap(Age ~ .) +
  ylab("Proportion (%)") +
  xlab("Satisfaction") +
  scale_fill_brewer(name = "Satisfaction", palette = 'Spectral', na.value = "gray") +
  ggtitle("Overall Satisfaction and Age")
```



From the chart, we can see that younger respondants more frequently give positive ratings, while as age increases there are proportionally more negative reviews.

2. Gender:

From the histogram, more males completed the surveys than females. Both rated the airport as 4 the most.

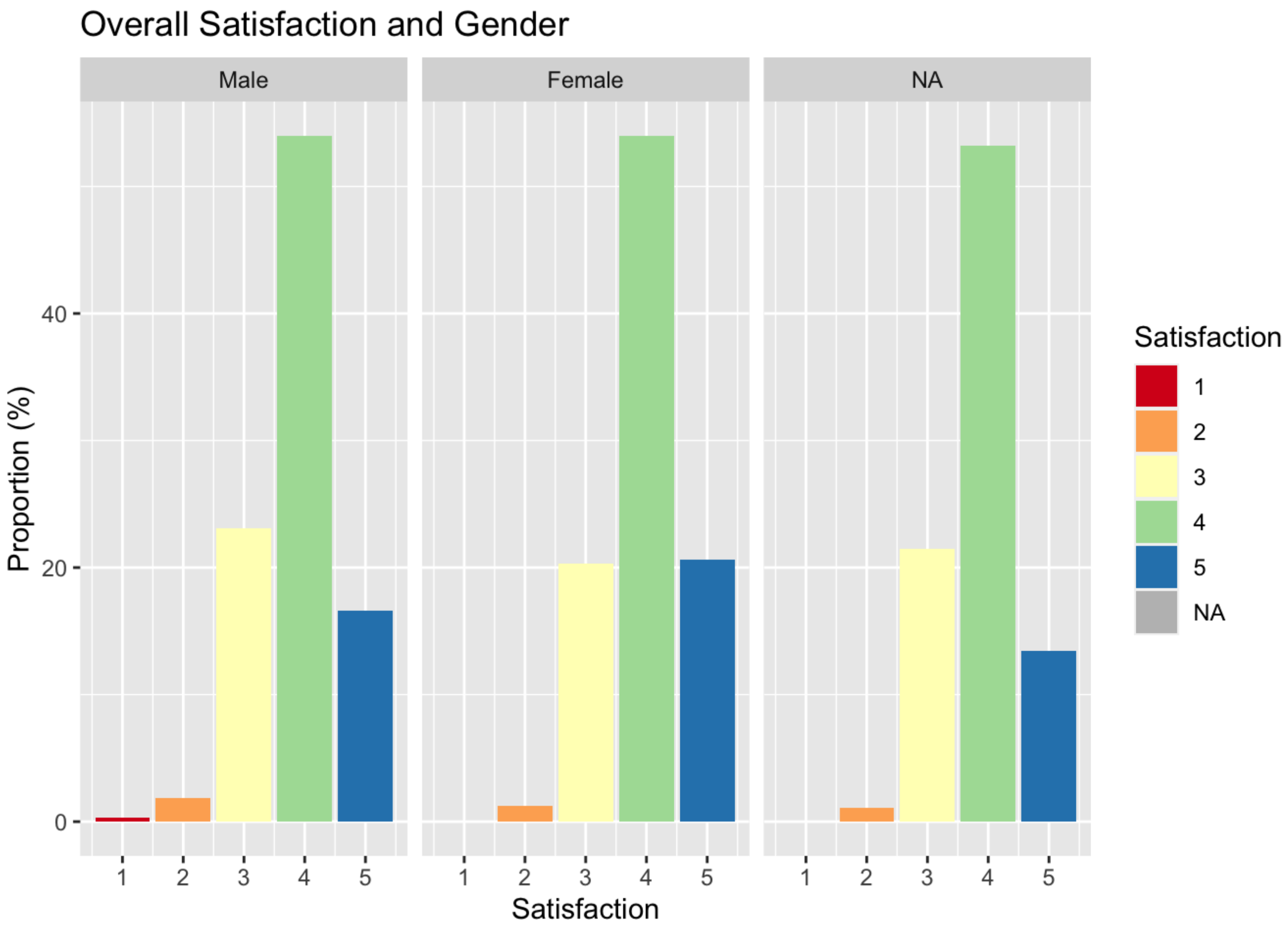
```
# Give names to Gender
levels(q1_factors$Gender) <- c("Male","Female")

q1_prop <- q1_factors %>%
  group_by(Gender) %>%
  summarise(n_gender = n()) %>%
  left_join(q1_factors %>%
            group_by(Gender, CFO_Airport_as_a_whole_0_6) %>%
            summarise(n_sd_inside_sat = n())) %>%
  mutate(gender_prop = (n_sd_inside_sat / n_gender) * 100)
```

```
## `summarise()` has grouped output by 'Gender'. You can override using the `.groups` ar
gument.
```

```
## Joining, by = "Gender"
```

```
ggplot(data = q1_prop, aes(x= CFO_Airport_as_a_whole_0_6, y = gender_prop, fill = as.fac
tor(CFO_Airport_as_a_whole_0_6))) +
  geom_col(position = 'dodge') +
  facet_wrap(Gender ~ .) +
  ylab("Proportion (%)") +
  xlab("Satisfaction") +
  scale_fill_brewer(name = "Satisfaction", palette = 'Spectral', na.value = "gray") +
  ggtitle("Overall Satisfaction and Gender")
```

```
## Warning: Removed 3 rows containing missing values (geom_col).
```



While a rating of 4 is most frequent across all genders, males appear to be more likley than females to give lower satisfaction scores.

3. Income:

From the histogram, those with an income of $50,000 to $100,000 had the highest response rate of the survey. Most of the income groups rated the airport as 4 the most.

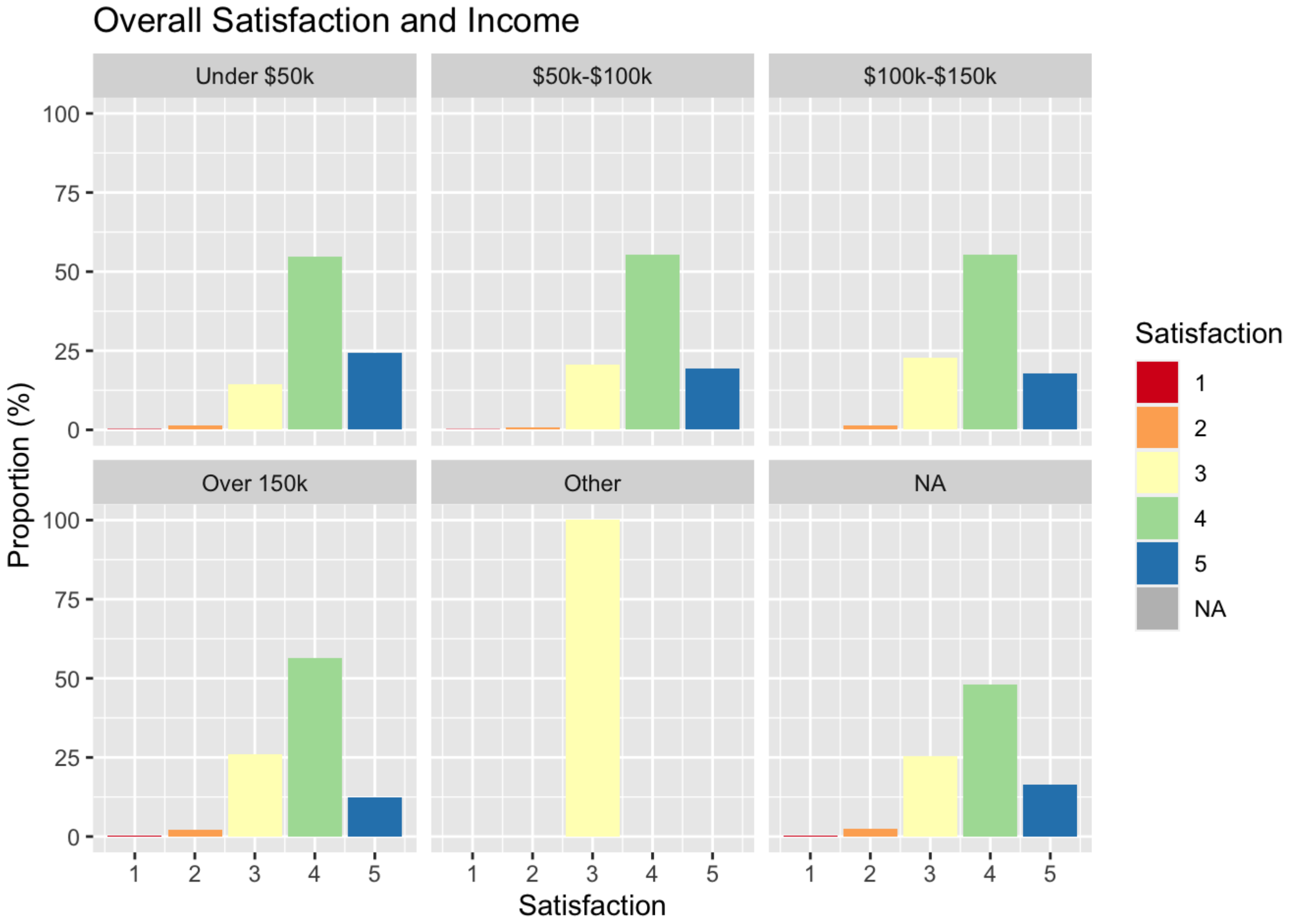
```
# Give names to Income
levels(q1_factors$Income_code) <- c("Under $50k","$50k-$100k","$100k-$150k","Over 150k",
"Other")

q1_prop <- q1_factors %>%
  group_by(Income_code) %>%
  summarise(n_income = n()) %>%
  left_join(q1_factors %>%
              group_by(Income_code, CFO_Airport_as_a_whole_0_6) %>%
              summarise(n_sd_inside_sat = n())) %>%
  mutate(income_prop = (n_sd_inside_sat / n_income) * 100)
```

```
## `summarise()` has grouped output by 'Income_code'. You can override using the `.group
s` argument.
```

```
## Joining, by = "Income_code"
```

```
ggplot(data = q1_prop, aes(x= CFO_Airport_as_a_whole_0_6, y = income_prop, fill = as.fac
tor(CFO_Airport_as_a_whole_0_6))) +
  geom_col(position = 'dodge') +
  facet_wrap(Income_code ~ .) +
  ylab("Proportion (%)") +
  xlab("Satisfaction") +
  scale_fill_brewer(name = "Satisfaction", palette = 'Spectral', na.value = "gray") +
  ggtitle("Overall Satisfaction and Income")
```

```
## Warning: Removed 5 rows containing missing values (geom_col).
```



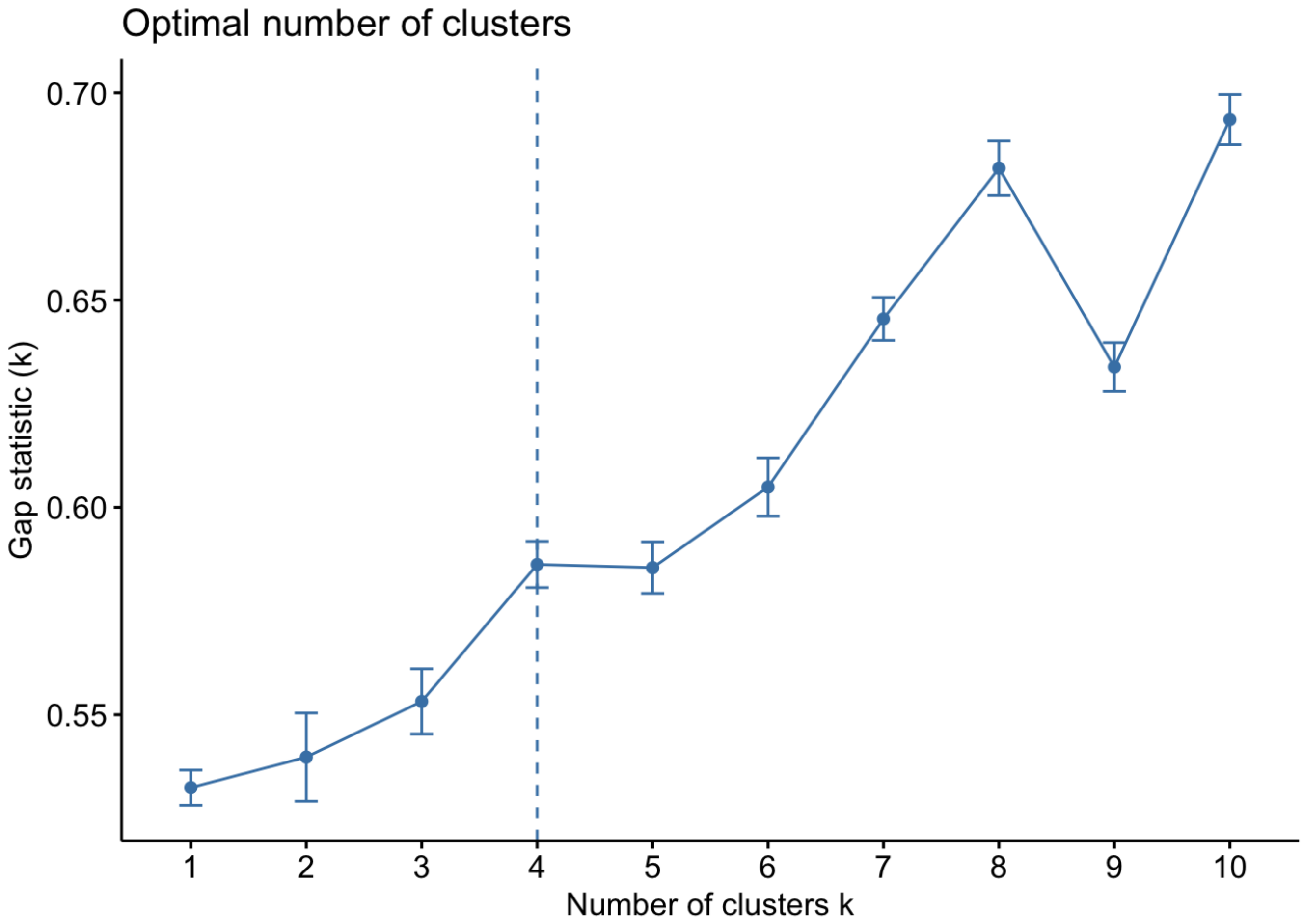
It appears that proportionally, those with higher income may be slightly less likley to be highly satisfied.

Next, we identify any patterns across all demographics.

We must determine how many clusters to use. This number is the number of groups that the different objects will be grouped in. We see from the graph below to use 4 groups.

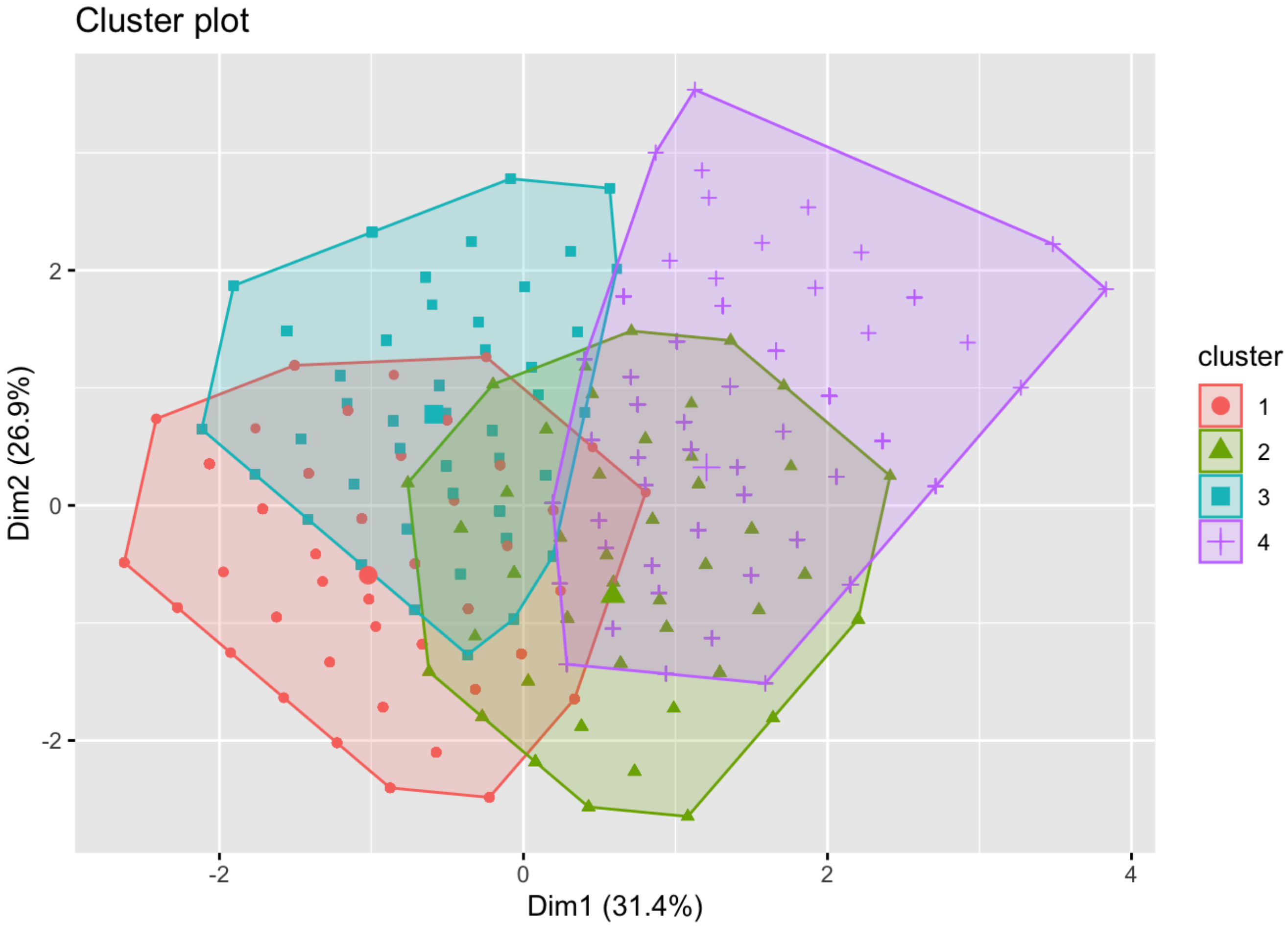
```
q1 %>%
  scale() %>%
  na.omit() %>%
  fviz_nbclust(x = ., FUNcluster = kmeans, method = "gap")
```

```
## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations
```



We used the kmeans test, where the different groups are formed by the closest observations to the mean of the cluster. Though this test is sensitive to outliers, since the data is based on 5 discrete responses, the presence of strong outliers is unlikely.

```
kmeansTest <-  kmeans(x = scale(na.omit(q1)), centers = 4)

fviz_cluster(kmeansTest, scale(na.omit(q1)), geom = "point")
```



We can view the values of the centers of the clusters.

```
kmeansTest$centers
```

```
##          Age Income_code     Gender CFO_Airport_as_a_whole_0_6
## 1 -0.1711538  -0.7830930  1.0969615                  0.2471944
## 2  0.2777765   0.9240385  1.0969615                 -0.1615855
## 3 -0.4974957  -0.7021737 -0.9112326                  0.1531717
## 4  0.4889847   0.8479368 -0.9112326                 -0.2885583
```

However, these centers are scaled. We look at the unscaled centers for further analysis.

```
usedData <-  q1 %>%
  na.omit() %>%
  scale()

kmeansCenters <- as.data.frame(kmeansTest$centers)

varNames <-  names(kmeansCenters)

unscaleCenters <-  function(varName) {
  unscaled = kmeansCenters[, varName] *
    attr(x = usedData, which = "scaled:scale")[varName] +
    attr(x = usedData, which = "scaled:center")[varName]

  res = data.frame(unscaled)

  names(res) = varName

  return(res)
}

clusterCentersRaw <-  lapply(varNames, function(x) unscaleCenters(x)) %>%
  bind_cols(.)
clusterCentersRaw %>%
 mutate_if(is.numeric, round, digits = 2)
```

```
##    Age Income_code Gender CFO_Airport_as_a_whole_0_6
## 1 4.10        1.58      1                       4.12
## 2 4.77        3.44      1                       3.83
## 3 3.60        1.66      0                       4.05
## 4 5.09        3.35      0                       3.74
```

We see that the clusters demonstrate that there is a pattern across satisfied and dissatisfied customers based on demographic characteristics.
Cluster 1: 35-44, Around 75k, female, highly positive Cluster 2: 44-55, Around 150k, female, satisfactory Cluster 3: 25-35, Around 75k, male, highly positive Cluster 4: 45-54, Around 150k, male, satisfactory

Question 2

The executives want to know if customer satisfaction can be broken down into different attributes of the airport.

To determine how customers rated each individual aspect of the airport, we can plot the airport aspect against the customer satisfaction. To determine if there are any broad themes, we used factor analysis.

Factor analysis leverages covariance or correlation among the variables to detect larger, underlying variables with more meaning and weight than the originally measured variables. The factors we detect represent some underlying construct of satisfaction. These factors are unobservable variables but reduce the originally measured responses into a smaller, more actionable set of factors.

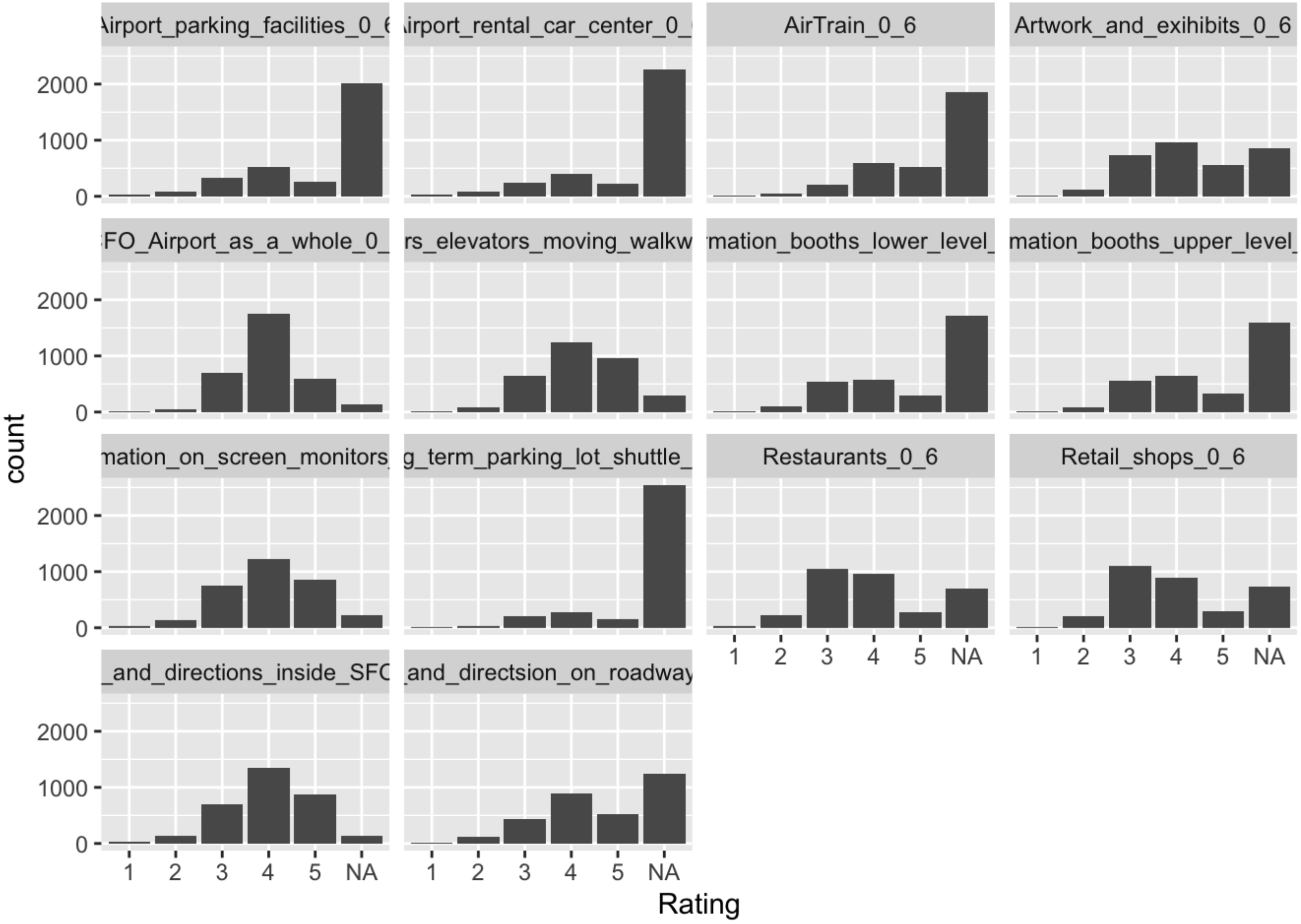
Created subset with overall customer satisfaction and satisfaction of the other airport aspects.

```
q2 <- missing_encoded %>%
  dplyr::select(Artwork_and_exihibits_0_6,
         Restaurants_0_6,
         Retail_shops_0_6,
         Signs_and_directions_inside_SFO_0_6,
         Escalators_elevators_moving_walkways_0_6,
         Information_on_screen_monitors_0_6,
         Information_booths_lower_level_0_6,
         Information_booths_upper_level_0_6,
         Signs_and_directsion_on_roadways_0_6,
         Airport_parking_facilities_0_6,
         AirTrain_0_6,
         Long_term_parking_lot_shuttle_0_6,
         Airport_rental_car_center_0_6,
         CFO_Airport_as_a_whole_0_6)
```

The customer satisfaction by each individual airport aspect can be seen as the following. We see that the highest ratings are 3s and 4s for most of the airport aspects.

```
q2 %>% pivot_longer(cols = ends_with("0_6")) %>%
#na.omit() %>%
mutate(value = ifelse(value == 6, NA, value)) %>%
ggplot(aes(x=as.factor(value))) +
geom_histogram(stat = 'count') +
facet_wrap(~name, nrow = 4) +
xlab('Rating')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



We then perform factor analysis to identify and analyze the latent variables. However, we first determine if factor analysis can be performed on this data set.

```
cortest.bartlett(q2)
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 30855.14
##
## $p.value
## [1] 0
##
## $df
## [1] 91
```

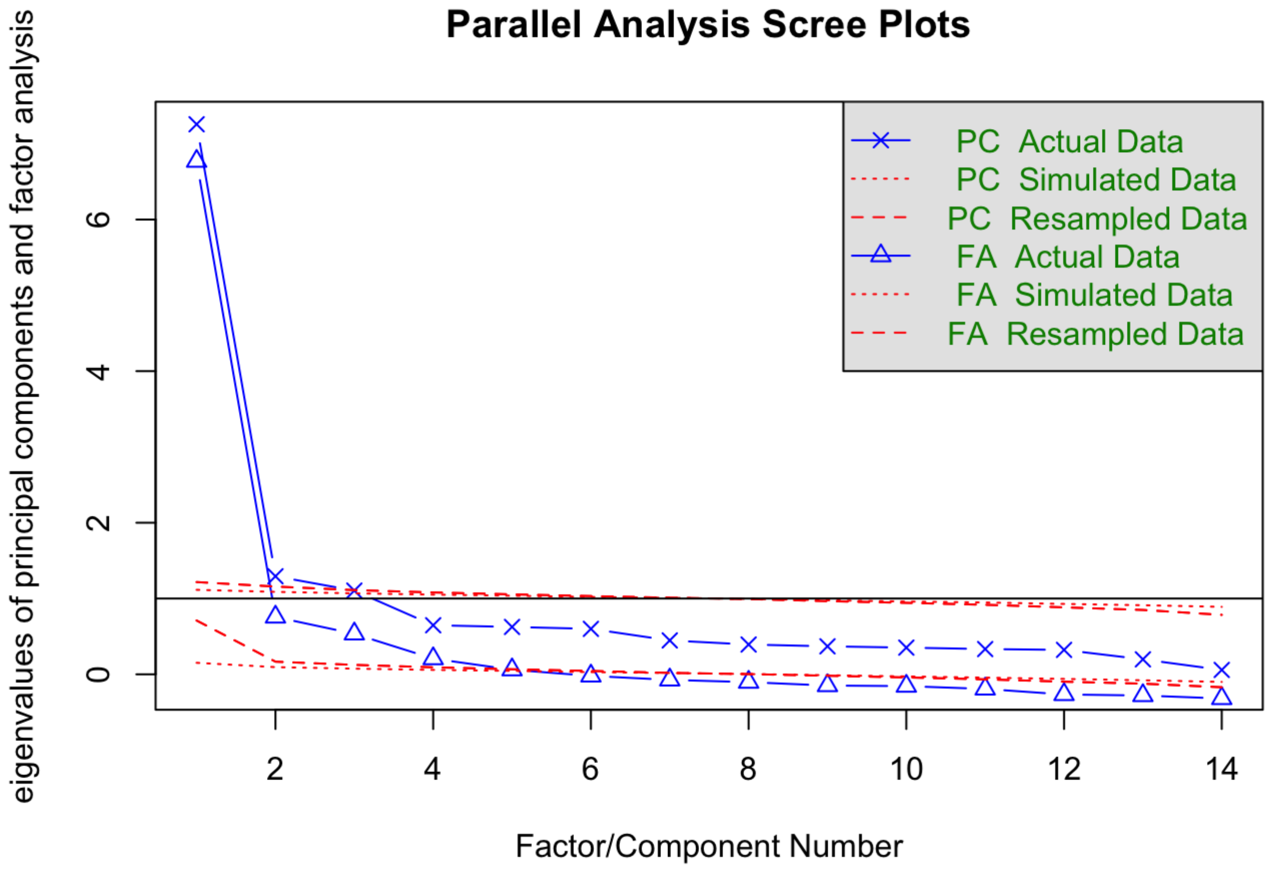
```
KMO(q2)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = q2)
## Overall MSA =  0.91
## MSA for each item = 
##                Artwork_and_exihibits_0_6 
##                                     0.92 
##                          Restaurants_0_6 
##                                     0.88 
##                         Retail_shops_0_6 
##                                     0.90 
##      Signs_and_directions_inside_SFO_0_6 
##                                     0.95 
## Escalators_elevators_moving_walkways_0_6 
##                                     0.95 
##       Information_on_screen_monitors_0_6 
##                                     0.96 
##       Information_booths_lower_level_0_6 
##                                     0.84 
##       Information_booths_upper_level_0_6 
##                                     0.84 
##     Signs_and_directsion_on_roadways_0_6 
##                                     0.95 
##           Airport_parking_facilities_0_6 
##                                     0.94 
##                             AirTrain_0_6 
##                                     0.93 
##        Long_term_parking_lot_shuttle_0_6 
##                                     0.89 
##            Airport_rental_car_center_0_6 
##                                     0.90 
##               CFO_Airport_as_a_whole_0_6 
##                                     0.95
```

We see that the p value from the barlett test is very small, so we reject the null hypothesis that the variables are not correlated. The overall MSA is greater that 0.6. Therefore, it makes sense to perform factor analysis to combine the variables.

We determine the optimal number of factors that explain variation among all the variables, while reducing the number of dimensions.

```
psych::fa.parallel(q2)
```



```
## Parallel analysis suggests that the number of factors =  5  and the number of compone
nts =  3
```

This scree plot shows when additional factors no longer add additional explanatory value. The threshold is where the FA Actual Data line drops below the simulated or resampled data. In this case it is after 4 factors. However, the line is so close to the simulated line, that we leaned on parsimony and used three factors in our final analysis.

We then fit the model with oblique rotation allowing for factors that contain some correlation.

```
x3 <- fa(r = q2, nfactors = 3, rotate = "promax")
```

```
## Loading required namespace: GPArotation
```

```
x3$loadings
```

```
## 
## Loadings:
##                                          MR1    MR2    MR3   
## Artwork_and_exihibits_0_6                 0.171         0.484
## Restaurants_0_6                          -0.132         0.901
## Retail_shops_0_6                                        0.792
## Signs_and_directions_inside_SFO_0_6       0.627  0.130       
## Escalators_elevators_moving_walkways_0_6  0.654              
## Information_on_screen_monitors_0_6        0.857              
## Information_booths_lower_level_0_6        0.917              
## Information_booths_upper_level_0_6        0.907              
## Signs_and_directsion_on_roadways_0_6      0.210  0.515       
## Airport_parking_facilities_0_6                   0.792       
## AirTrain_0_6                                     0.770       
## Long_term_parking_lot_shuttle_0_6                0.903       
## Airport_rental_car_center_0_6                    0.792       
## CFO_Airport_as_a_whole_0_6                0.290  0.265  0.356
## 
##                  MR1   MR2   MR3
## SS loadings    3.397 3.030 1.808
## Proportion Var 0.243 0.216 0.129
## Cumulative Var 0.243 0.459 0.588
```

The first factor is related to restaurants, artwork, signs, escalations, and information booth. This could be associated with satisfaction with signs and visual information.

The second factor is related to roadways, parking, rental car, and airtrain. This could be associated with satisfaction with parking and transportation.

The last factor relates to artwork, restaurants, and retail. This could be associated with the retail aspects of the airport.

Question 3

Free-response comments, either positive or negative, were collected in addition to the 14-item quantitative survey. We can leverage these open ended responses to mine for sentiment within each response, and for overall topics that may or may not have been included in the survey structure.

We will first develop a sentiment analysis model to determines the overall sentiment of SFO guests by comparing positive and negative words. Since the comments are text based fields, we can look at the specific model of the comments to determine the overall sentiment.

We will then perform a topic analysis to see which specific topics were mentioned by the customers of the airport.

Created subset with customer satisfaction and comments

```
q3 <- SFOdata %>% dplyr::select(RESPNUM,Q7_text_All)
```

We see that there are more negative comments that positive comments in our sentiment model.

```
comments <- q3 %>%
  unnest_tokens(word, Q7_text_All)
# comments  %>%
#     inner_join(get_sentiments("nrc")) %>%
#     count(sentiment, sort = TRUE)
comments %>%
      inner_join(get_sentiments("bing")) %>%
      count(sentiment, sort = TRUE)
```

```
##   sentiment    n
## 1  negative 2079
## 2  positive 1594
```

Overall this is not a good result. Depending on how these responses are collected there may be a bias for dissatisfied customers to respond to the survey. We also do no have a baseline for whether airports in general perform very poorly in these sort of surveys, and this is actually a relatively good result, or vice-versa. Regardless, the sentiment here is overall quite negative.

Topic Model

For more specific analysis, we will develop a topic model. Topics models define the topics within each response, also referred to as a document. Combining all the comments regarding customers' overall satisfaction, our analysis determines the topics.

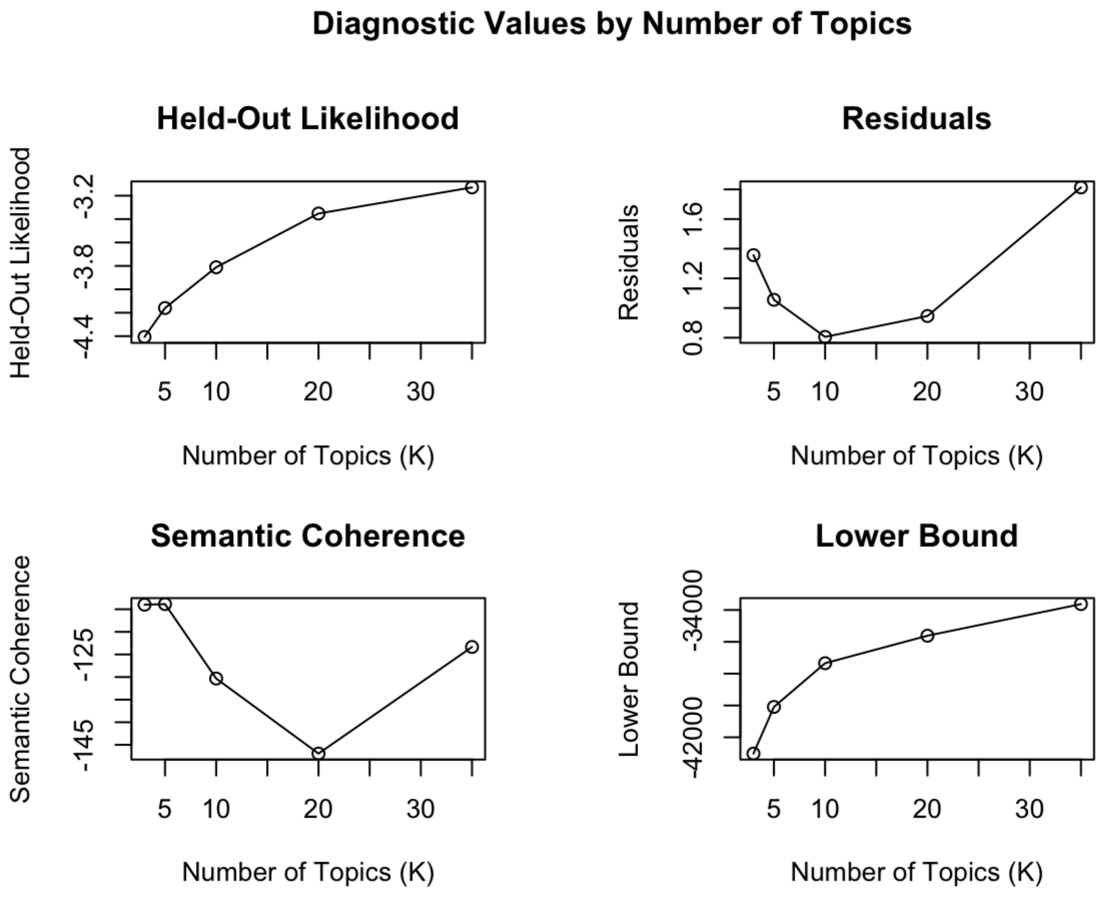
```
set.seed(1842)
custom_stop_words <- c("General",
                       "positive",
                       "negative",
                       "comments",
                       "about",
                       "SFO",
                       "sfo",
                       "airport",
                       "Airlines",
                       "neg",
                       "comment")

df_text <- textProcessor(documents = q3$Q7_text_All,
                         metadata = missing_encoded,
                         onlycharacter = TRUE,
                         customstopwords = c(tm::stopwords("SMART"),
                                             tm::stopwords("en"),
                                             custom_stop_words))
```

```
## Building corpus...
## Converting to Lower Case...
## Removing punctuation...
## Removing stopwords...
## Remove Custom Stopwords...
## Removing numbers...
## Stemming...
## Creating Output...
```

```
df_Prep <-  prepDocuments(documents = df_text$documents,
                               vocab = df_text$vocab,
                               meta = df_text$meta)
```

```
## Removing 5 of 199 terms (5 of 9563 tokens) due to frequency
## Your corpus now has 1557 documents, 194 terms and 9558 tokens.
```

```
kTest <- searchK(documents = df_Prep$documents,
             vocab = df_Prep$vocab,
             K = c(3, 5, 10, 20, 35),
             verbose = F,
             max.em.its = 200)

plot(kTest)
```

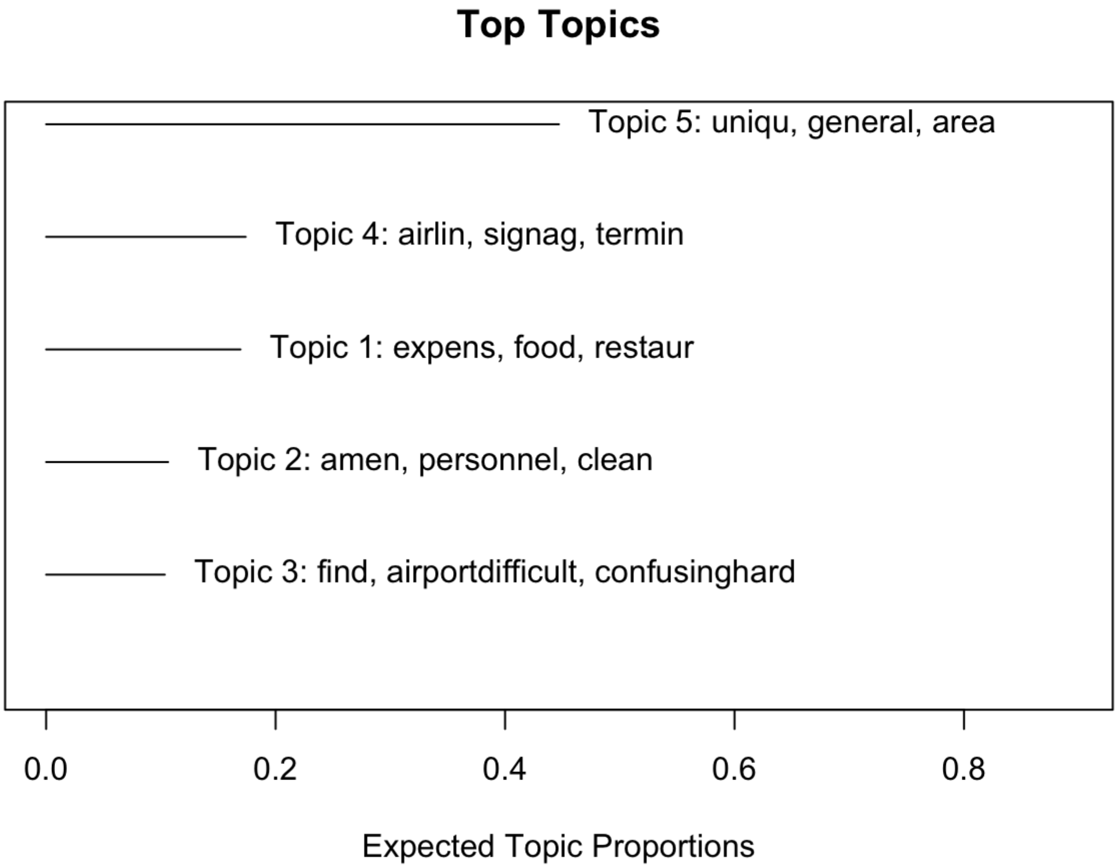


We decided to choose K = 5 due to the high semantic coherence (how well the words hang together) and low residuals. Even though K = 10 had the lowest residuals, we believed it may lead to high variance and overfit the data by being overly complex.

```
topics <-  stm(documents = df_Prep$documents,
             vocab = df_Prep$vocab,
             K = 5, #change this to whatever the plot shows best
             verbose = F,
             max.em.its = 200)
```

The plot and the labels represent the different topics from the comments.

```
plot(topics)
```



```
labelTopics(topics)
```

```
## Topic 1 Top Words:
##       Highest Prob: expens, food, restaur, secur, healthier, low, qualityne
##       FREX: food, secur, healthier, low, qualityne, select, crowdedmor
##       Lift: enoughne, healthier, low, move, qualityne, restaurantsstoresclub, select
##       Score: food, healthier, low, qualityne, select, crowdedmor, expens
## Topic 2 Top Words:
##       Highest Prob: amen, personnel, clean, hook, miss, small, restroom
##       FREX: amen, personnel, clean, hook, miss, small, check
##       Lift: atm, check, clock, curbsid, hand, payphon, sanit
##       Score: amen, clean, hook, miss, small, payphon, sanit
## Topic 3 Top Words:
##       Highest Prob: find, airportdifficult, confusinghard, correct, awaydifficult, ca
r, center
##       FREX: airportdifficult, confusinghard, correct, awaydifficult, car, center, ren
tal
##       Lift: electronicautom, enoughinconveni, humansfew, locatedshould, luggag, sign,
airportdifficult
##       Score: bilingu, airportdifficult, confusinghard, correct, rental, rude, toconfu
singemploye
## Topic 4 Top Words:
##       Highest Prob: airlin, signag, termin, find, confusingsmallhard, gate, insid
##       FREX: airlin, confusingsmallhard, gate, insid, airportno, awayconfusingtoo, dif
ficulttak
##       Lift: confusingsmallhard, gate, insid, add, airlin, airportno, awayconfusingtoo
##       Score: airlin, signag, termin, confusingsmallhard, gate, insid, signsannounceme
ntspersonnel
## Topic 5 Top Words:
##       Highest Prob: uniqu, general, area, inform, securitycustom, linesprocedur, long
inefficientineffect
##       FREX: general, area, inform, linesprocedur, longinefficientineffect, chang, eno
ughlack
##       Lift: air, allow, artwork, artworkexhibitionschang, control, cut, delaysbad
##       Score: area, electr, uniqu, inform, general, linesprocedur, longinefficientinef
fect
```

From our Topic Model, we see the topics from the comments are separated into the below 5 groups in order of frequency:

Topic 5: Difficulty moving through security Topic 4: Confusion around signage and navigating through the airport Topic 1: Issues with quality, price, and selection of restaurants Topic 2: Cleanliness and sanitation Topic 3: Confusion around car rentals

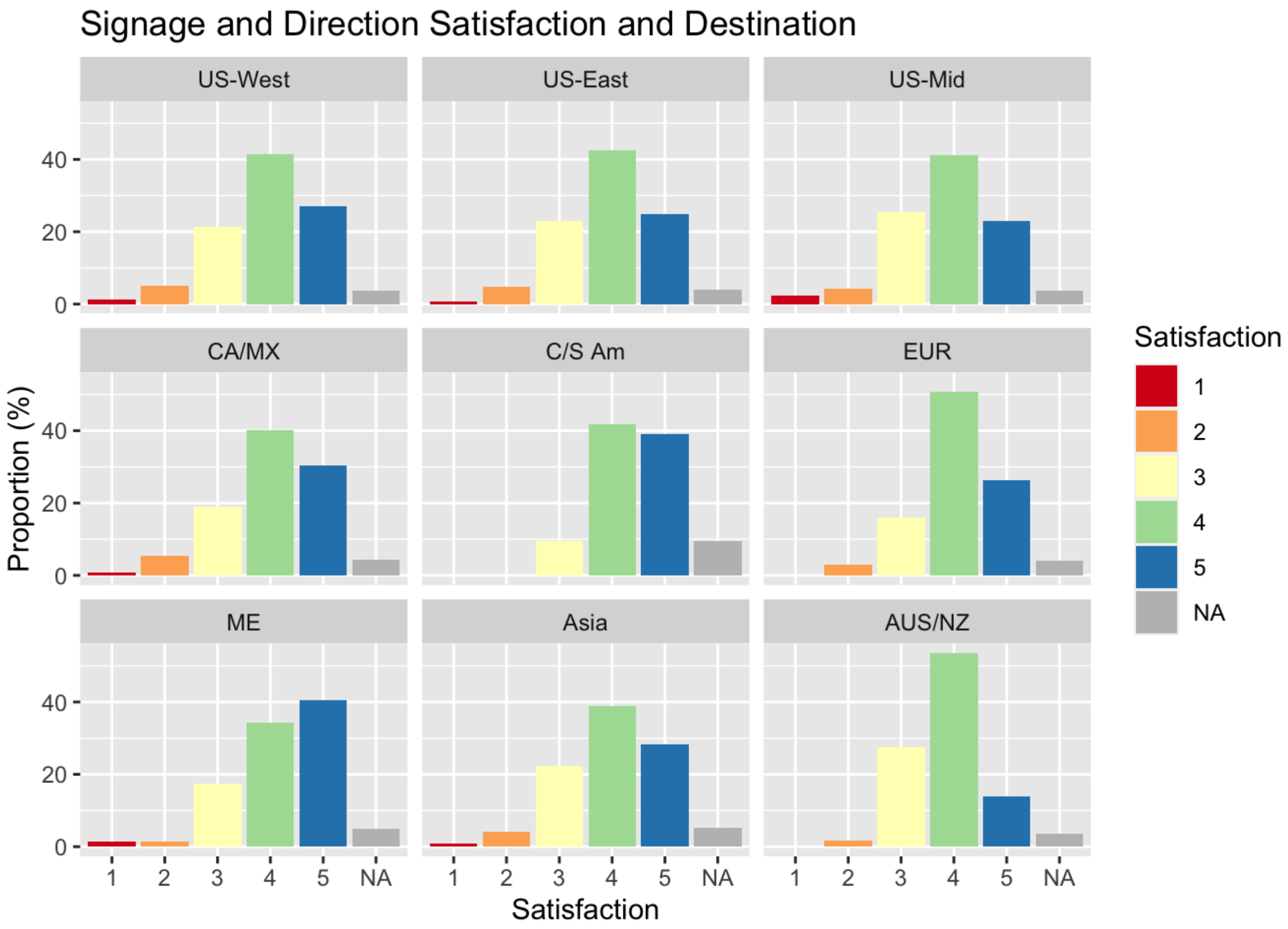
Most of these topics focus on negative aspects of the airport, with the most frequent being comments around difficult navigation through the airport whether it be security or signage.

Part B

The question we tackle in this section is ‘Do particular groups of passengers have more trouble with signage (one of the most commented issue areas) than others?’ i.e can accessibility be improved by identifying and targeting improvements for at risk groups. Possible variables to investigate: “Age”, “Gender”, “Income_code”, “Destination_Geo_Loc”.

```
## `summarise()` has grouped output by 'Destination_Geo_Loc'. You can override using the
`.groups` argument.
```

```
## Joining, by = "Destination_Geo_Loc"
```



Using the location of the destination airport, we can see that responders flying inside the US have a high proportion of neutral to negative responses compared to other destinations. This may be a useful variable for targeting signage changes for passengers flying domestic.

LCA is used with discrete data, such as survey responses since it only works with categorical data. Since the dataset is not made up of one homogenous group, it is better modeled by multiple groups that are characterized by unique response propensities. We then review profiles for the different groups. We decided to use LCA because it does a good job of grouping responders together in insightful ways. This will hopefully allow us to determine how we can target specific groups which may need more help with signage than others.

```
#Create the responses as factors
part_b <-  missing_encoded %>%
  dplyr::select(Age,
         Gender,
         Income_code,
         Destination_Geo_Loc, # destination US region or foreign continent
         Signs_and_directions_inside_SFO_0_6,
         Signs_and_directsion_on_roadways_0_6) %>%
  mutate(Destination_Geo_Loc = as.factor(Destination_Geo_Loc))
part_b$Age <- sapply(part_b$Age, function(x) (as.numeric(factor(x,levels = c(1,2,3,4,5,6
,7)))))
part_b$Gender <- sapply(part_b$Gender, function(x) (as.numeric(factor(x,levels = c(0,1
)))))
part_b$Income_code <- sapply(part_b$Income_code, function(x) (as.numeric(factor(x,levels
= c(1,2,3,4,5)))))
part_b$Signs_and_directions_inside_SFO_0_6 <- sapply(part_b$Signs_and_directions_inside_
SFO_0_6, function(x) (as.numeric(factor(x,levels = c(1,2,3,4,5)))))
part_b$Signs_and_directsion_on_roadways_0_6 <- sapply(part_b$Signs_and_directsion_on_roa
dways_0_6, function(x) (as.numeric(factor(x,levels = c(1,2,3,4,5)))))
part_b$Destination_Geo_Loc <- sapply(part_b$Destination_Geo_Loc, function(x) (as.numeric
(factor(x,levels = c(1,2,3,4,5,6,7,8,9,10)))))

part_b <- part_b %>%
  rename(`Signs Inside` = `Signs_and_directions_inside_SFO_0_6`,
         `Signs Outside` = `Signs_and_directsion_on_roadways_0_6`)
```

We create different models with different number of classes. We then determine which model is the best to use.

To determine which is the best model, we compare AIC for statistical criterion.

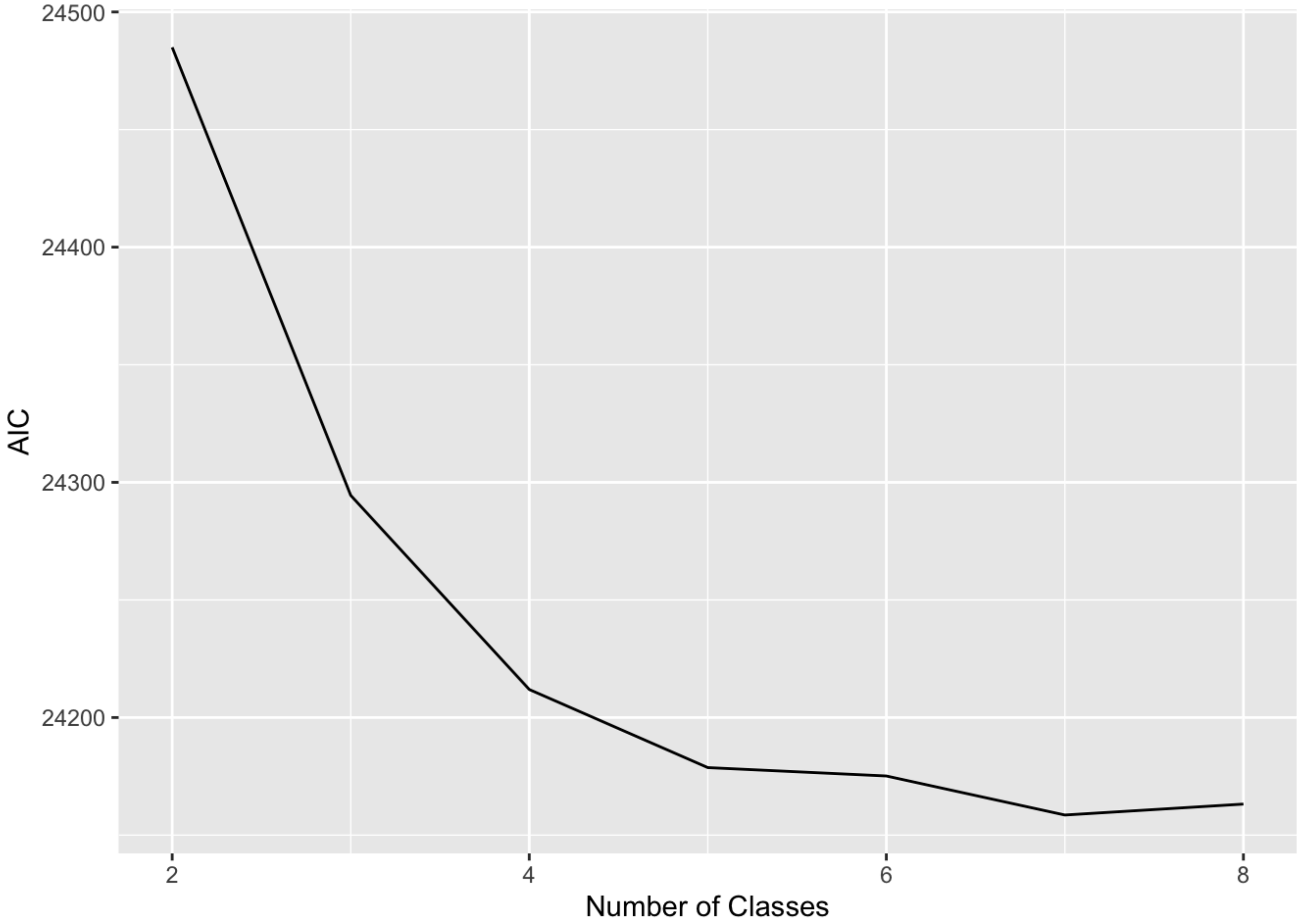
```
rbind(class2 = lcaMod2$aic,
      class3 = lcaMod3$aic,
      class4 = lcaMod4$aic,
      class5 = lcaMod5$aic,
      class6 = lcaMod6$aic,
      class7 = lcaMod7$aic,
      class8 = lcaMod8$aic)
```

```
##            [,1]
## class2 24484.96
## class3 24294.46
## class4 24211.94
## class5 24178.71
## class6 24175.15
## class7 24158.54
## class8 24163.20
```

```
class_aic <- c(lcaMod2$aic,
  lcaMod3$aic,
  lcaMod4$aic,
  lcaMod5$aic,
  lcaMod6$aic,
  lcaMod7$aic,
  lcaMod8$aic
  ) %>%
  cbind(2:8) %>%
  as_tibble()
```

```
## Warning: The `x` argument of `as_tibble.matrix()` must have unique column names if `.
name_repair` is omitted as of tibble 2.0.0.
## Using compatibility `.name_repair`.
```

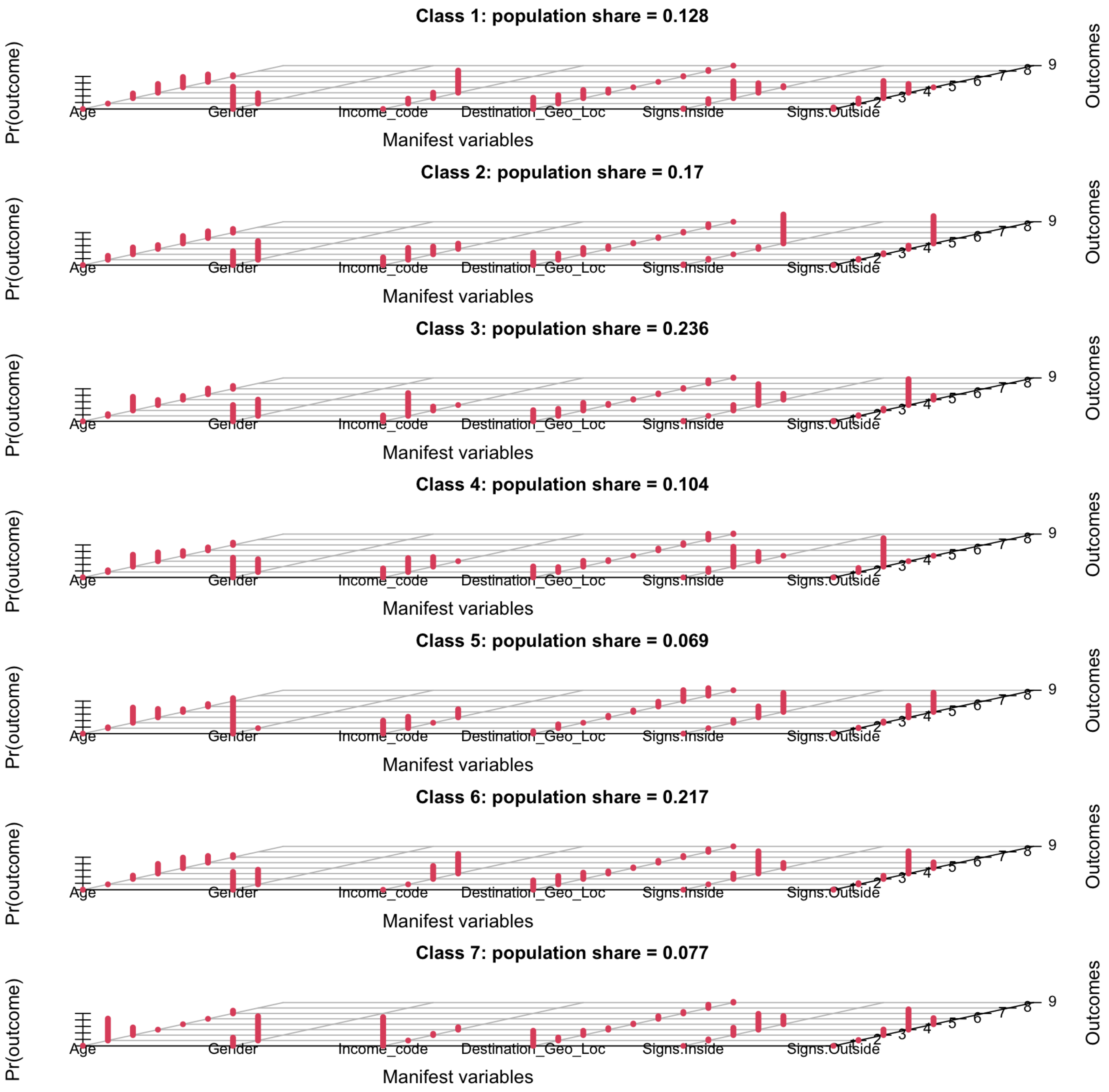
```
ggplot(data = class_aic,aes(y = ., x= V2))+
  geom_line()+
  labs(x = 'Number of Classes', y = 'AIC')
```



The AIC drops until there are seven classes. This suggests that the best model for explaining the variability between responses is to include 7 classes in the latent class model. There are 7 distinct types of customers differentiated by Age, Gender, Income, Destination and their ability to find signs and directions.

The plot below visualizes the seven different profiles from the customers.

```
plot(lcaMod7)
```



The seven classes of customers are the following build out three distinct profiles from the customers in relation to signage:

Class 1: Around the age of 35-44, with a salary between $50k-$150k, travelings within the United States, finds the signs and directions in SFO are very good, and the signage on SFO roadways as very good. Class 2: Adults of all ages and incomes traveling within the United States find the signs and directions in SFO as very good with the signage of SFO roadways as very good. Class 3:Middle age adults with a mid salary traveling within the United States find the signs and directions in SFO as satisfactory with the signage of SFO roadways as satisfactory. Class 4: Middle age adults with a lower to mid salary traveling within the United States find the signs and directions in SFO as fair/satisfactory with the signage of SFO roadways as fair/satisfactory. Class 5: Middle aged male adults with lower salary traveling internationally find the signs and directions in SFO as very good with the signage of SFO roadways as very positive. Class 6: Middle to older male adults with higher salary traveling within the United States find the signs and directions in SFO as very good with the signage of SFO roadways as very positive. Class 7: Young adults (age 18-34) with a a lower salary (under $50k) traveling within the United States find the signs and directions in SFO as very good with the signage of SFO roadways as satisfactory.

Based on our results, we found that responders from Class 4 (Middle age adults with a lower to mid salary traveling within the United States) have lower satisfaction with directions and signage, and therefore should be targeted for improvements.