Customer Satisfaction Analysis Report

AUTHOR
Cristian T

PUBLISHED
October 27, 2024

Introduction

This report presents an in-depth analysis of customer satisfaction based on the San Francisco International Airport (SFO) 2010 Customer Survey. With a set of questions addressing airport services and operations—including satisfaction with amenities, parking, information, and transportation—the survey aims to capture passenger perceptions across key service areas. Given the vital role of customer satisfaction in shaping passenger experiences and fostering loyalty, this analysis identifies and evaluates the factors that contribute to customer contentment or dissatisfaction at SFO.

Using Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA), the report uncovers underlying themes in customer satisfaction. EFA reveals the latent factors that structure customer responses, while CFA validates these dimensions to ensure a consistent, reliable framework for understanding customer needs. Demographic comparisons further enhance the analysis by highlighting how distinct passenger groups (e.g., by age, gender, and income) perceive airport services differently, thus informing targeted improvements.

To complement the factor analysis, we used sentiment scoring, which was conducted to quantify the sentiment expressed in structured feedback on service improvements. The analysis uses aggregate sentiment scores to capture the general sentiment intensity for specific aspects, such as dining options, retail offerings, and airport layout. This approach provides nuanced insights into positive and negative sentiment trends to pinpoint areas of praise and opportunities for enhancement.

The primary objective of this analysis is to deliver actionable insights that SFO executives can use to enhance service quality, improve satisfaction, and build passenger loyalty. By understanding the attributes that most influence satisfaction, SFO can strategically enhance strengths and address areas needing improvement, ensuring a more positive and memorable experience for all passengers. This report lays a strong foundation for informed decision-making and strategic planning within the competitive landscape of airport services.

Exploratory Factor Analysis (EFA) on Passenger Satisfaction

Exploratory Factor Analysis (EFA) was conducted to identify the underlying dimensions of customer satisfaction regarding airport services at San Francisco International Airport (SFO). This analysis focused on passenger ratings of the following attributes:

- 6a. Artwork and exhibitions

- 6b. Restaurants
- 6c. Retail shops and concessions
- 6d. Signs and directions inside SFO
- 6e. Escalators, elevators, moving walkways
- 6f. Information on screens and monitors
- 6g. Information booths (lower level near baggage claim)
- 6h. Information booths (upper level – departure area)
- 6i. Signs and directions on SFO airport roadways
- 6j. Airport parking facilities
- 6k. AirTrain
- 6l. Long-term parking lot shuttle
- 6m. Airport rental car center
- 6n. SFO Airport as a whole

Passengers rated these attributes on a scale from 1 (Unacceptable) to 5 (Outstanding), with additional options for those who had never used or visited the facilities or considered the question not applicable.

Data Preparation

To facilitate this analysis, we utilized several R libraries, and began by loading the dataset and inspecting its structure to ensure that the data was ready for analysis. This initial step helped us understand the variables at play and set the stage for the subsequent exploratory factor analysis.

```
suppressMessages(library(tidyverse))
suppressMessages(library(skimr))
suppressMessages(library(ggplot2))
suppressMessages(library(stringr))
suppressMessages(library(psych))
suppressMessages(library(lavaan))
suppressMessages(library(semPlot))
suppressMessages(library(dplyr))
suppressMessages(library(sentimentr))
suppressMessages(library(caret))

dat <- read.table("/Users/cristian/Downloads/ND Data/SFO_survey_withText.txt", header = T

# Get an overview of Data Structure
# glimpse(dat)


# Select the 14 customer satisfaction questions
customer_satisfaction_data <- dat %>% dplyr::select(Q6A:Q6N)


# Apply skim function to view summary statistics
skim(customer_satisfaction_data)
```

Data summary

Name	customer_satisfaction_dat…
Number of rows	3234
Number of columns	14
________________________	
Column type frequency:	
numeric	14
________________________	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Q6A	103	0.97	4.33	1.21	1	3.00	4	5	6	
Q6B	81	0.97	3.98	1.25	1	3.00	4	5	6	
Q6C	106	0.97	3.99	1.25	1	3.00	4	5	6	
Q6D	79	0.98	3.96	0.93	1	3.00	4	5	6	
Q6E	74	0.98	4.18	0.94	1	4.00	4	5	6	
Q6F	79	0.98	4.00	0.97	1	3.00	4	5	6	
Q6G	113	0.97	4.87	1.32	1	4.00	6	6	6	
Q6H	115	0.96	4.81	1.30	1	4.00	5	6	6	
Q6I	144	0.96	4.65	1.24	1	4.00	5	6	6	
Q6J	182	0.94	5.10	1.25	1	4.00	6	6	6	
Q6K	190	0.94	5.16	1.08	1	4.00	6	6	6	
Q6L	214	0.93	5.49	1.03	1	6.00	6	6	6	
Q6M	216	0.93	5.26	1.20	1	4.25	6	6	6	
Q6N	112	0.97	3.94	0.73	1	4.00	4	4	6	

```
# Set a seed for reproducibility
set.seed(1842)

#Create a vector of row indices
row_indices <- 1:nrow(customer_satisfaction_data)

# Randomly sample 70% of the data for the training set
train_indices <- sample(row_indices, size = 0.7 * length(row_indices))

# Create the training and test sets
train_data <- customer_satisfaction_data[train_indices, ]
```

```r
test_data <- customer_satisfaction_data[-train_indices, ]

# Verify the split
head(train_data)
```

	Q6A	Q6B	Q6C	Q6D	Q6E	Q6F	Q6G	Q6H	Q6I
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
3077	5	5	5	5	5	5	5	5	5
718	5	3	3	4	4	4	3	3	3
911	3	4	3	5	3	5	3	3	4
2120	5	3	3	5	5	5	6	6	5
2940	4	6	4	4	5	4	6	6	6
952	3	6	6	6	3	3	6	3	6

6 rows | 1-10 of 15 columns

```r
head(test_data)
```

	Q6A	Q6B	Q6C	Q6D	Q6E	Q6F	Q6G	Q6H	Q6I
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
2	6	3	3	4	4	6	6	6	4
3	6	6	3	4	3	5	6	6	6
6	6	6	6	4	5	5	6	6	6
7	3	2	2	3	3	4	3	3	4
14	6	3	6	4	4	6	6	6	6
15	6	3	4	5	5	5	6	6	2

6 rows | 1-10 of 15 columns

```r
nrow(train_data) # Number of rows in training data
```

```
[1] 2263
```

```r
nrow(test_data) # Number of rows in test data
```

```
[1] 971
```

We extracted a subset of the dataset, focusing on the 14 questions related to customer satisfaction, labeled from Q6A to Q6N. We used the skim function to summarize the data, confirming that it contains 3,234 rows and 14 numeric columns. The summary revealed some missing values, with overall response completeness ranging from 93.3% to 97.7%. Overall, the data appeared to be in good shape for analysis.

To ensure the reproducibility of our analysis, we set a random seed, which enabled us to create a vector of all row indices. This vector was utilized to randomly sample 70% of the data for the training set, while the remaining 30% was allocated for testing. Consequently, the training dataset comprises 2,263 rows, while the test dataset contains 971 rows.

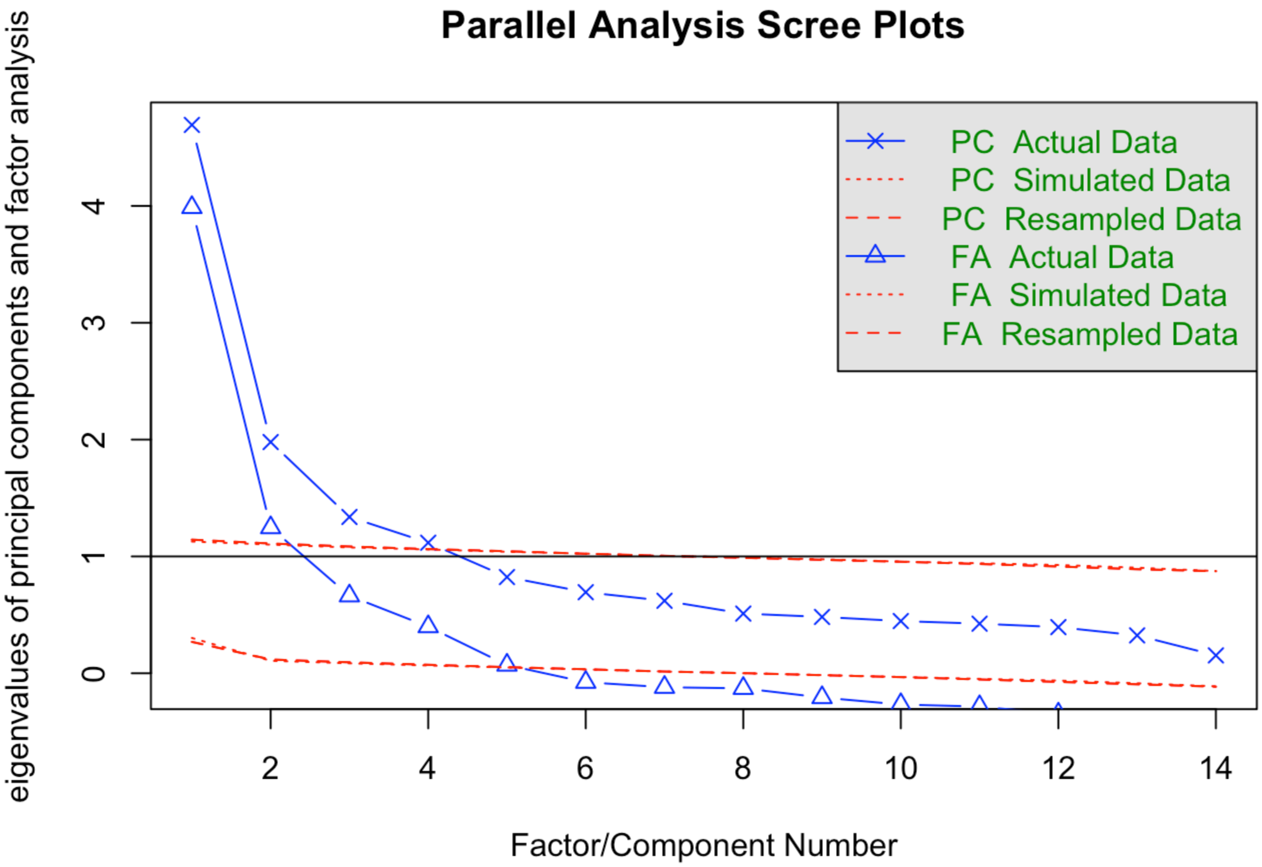
Kaiser-Meyer-Olkin Test and Factor Extraction

The analysis began with the Kaiser-Meyer-Olkin (KMO) test on the training dataset, yielding an overall Measure of Sampling Adequacy (MSA) of 0.82. Individual MSAs for the items ranged from 0.70 to 0.93, indicating that the data is suitable for factor analysis. Following this, a parallel analysis suggested extracting five factors, though manual counting indicated that four factors might be more appropriate. We conducted exploratory factor analysis (EFA) with 4 and 5 factors to compare the results and their implications for passenger satisfaction.

```r
# Use KMO to assess adequacy
KMO(train_data)
```

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = train_data)
Overall MSA =  0.82
MSA for each item =
 Q6A  Q6B  Q6C  Q6D  Q6E  Q6F  Q6G  Q6H  Q6I  Q6J  Q6K  Q6L  Q6M  Q6N
0.93 0.83 0.85 0.86 0.91 0.88 0.70 0.70 0.85 0.76 0.83 0.80 0.80 0.88
```

```r
# Determine the number of factors to extract
fa.parallel(train_data)
```




```
Parallel analysis suggests that the number of factors =  5  and the number of components
=  4
```

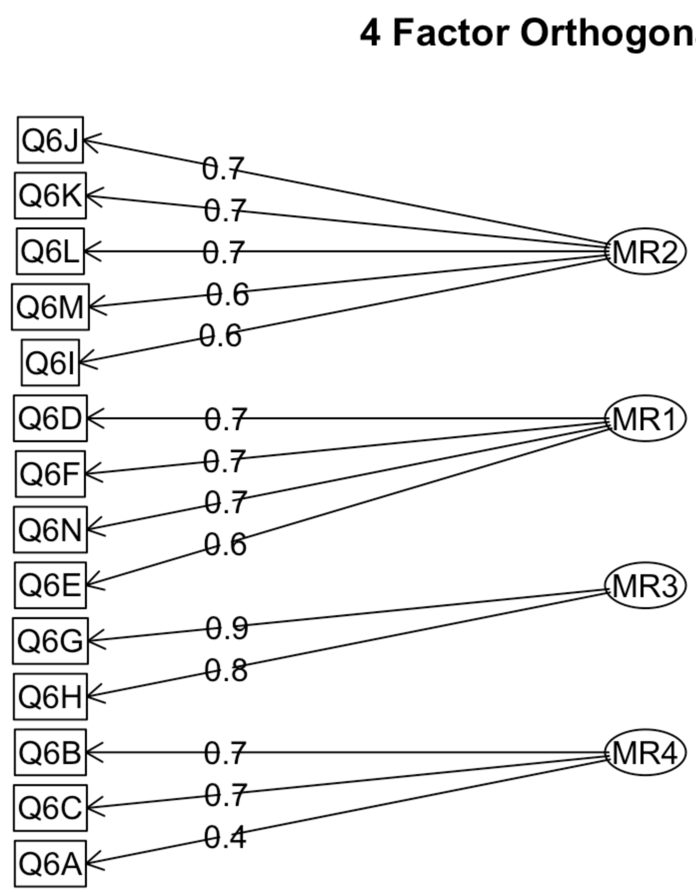
We explored two variations for each model: the Four-Factor Model and the Five-Factor Model with Varimax rotation, which aims to achieve orthogonality (making factors uncorrelated or independent from one another), as well as the Four-Factor Model and Five-Factor Model with Promax rotation, which allows for correlated factors.

```r
############## four-factor EFA model Varimax ##############
# Fit the four-factor FA model with orthogonal rotation (varimax):
train_FA_orth_4 <- fa(train_data, nfactors = 4, rotate = "varimax")

# Print the orthogonal rotation factor loadings
print(train_FA_orth_4$loadings, cutoff = 0.001, digits = 3, sort = TRUE)
```

```
Loadings:
    MR2   MR1   MR3   MR4
Q6I 0.620 0.181 0.092 0.110
Q6J 0.673 0.049 0.118 0.129
Q6K 0.670 0.142 0.046 0.110
Q6L 0.660 0.004 0.144 0.074
Q6M 0.625 0.115 0.060 0.096
Q6D 0.087 0.728 0.099 0.099
```

```
Q6E 0.180 0.613 0.175 0.238
Q6F 0.087 0.682 0.147 0.150
Q6N 0.122 0.657 0.045 0.278
Q6G 0.202 0.166 0.889 0.163
Q6H 0.172 0.205 0.844 0.143
Q6B 0.156 0.198 0.065 0.738
Q6C 0.176 0.243 0.112 0.653
Q6A 0.105 0.220 0.217 0.384

                 MR2   MR1   MR3   MR4
SS loadings    2.310 2.089 1.680 1.388
Proportion Var 0.165 0.149 0.120 0.099
Cumulative Var 0.165 0.314 0.434 0.533
```

```
# Display the plot
fa.diagram(train_FA_orth_4,
           main = "4 Factor Orthogonal (Varimax)")
```



Four-Factor Model with Varimax Rotation

The Four-Factor Model with Varimax Rotation provides a clear structure for understanding San Francisco International Airport (SFO) passenger satisfaction. Here, MR1 assesses the effectiveness of signage and information provided to passengers, highlighting attributes like signs and directions that help travelers navigate the airport. MR2 focuses on the convenience and efficiency of transportation options within the airport, including the AirTrain, parking facilities, and shuttle services. MR3 emphasizes the importance of information booths for assisting passengers. MR4 pertains to the quality and variety of food and shopping options available to passengers. These four factors explain 53.3% of the total variance in passenger satisfaction, indicating that more than half of the differences in passenger satisfaction can be attributed to these four factors.

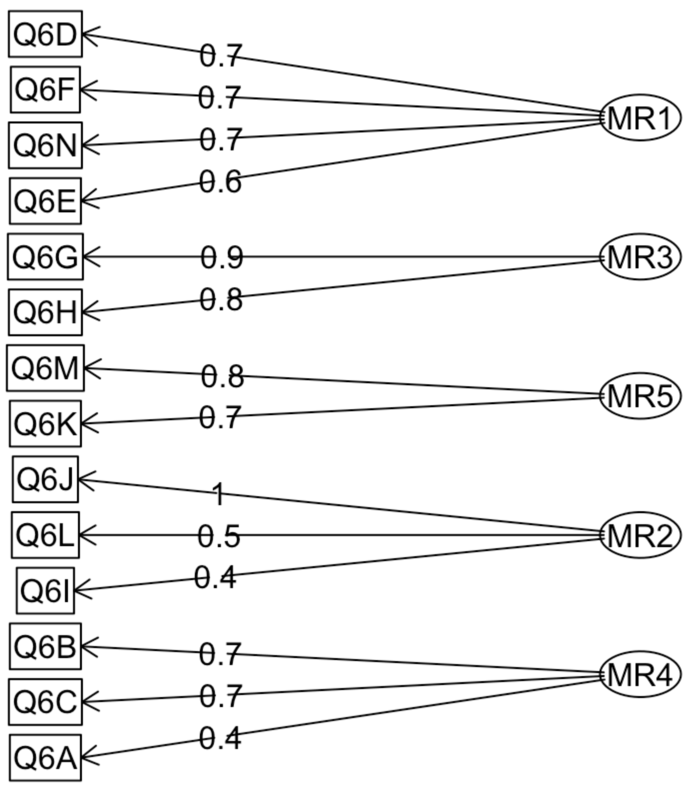
Transitioning to the Five-Factor Model with Varimax Rotation, we introduce an additional factor to enhance our understanding of passenger experiences.

```
############## five-factor EFA model Varimax ##############

# Fit the five-factor FA model with orthogonal rotation (varimax):
train_FA_orth <- fa(train_data, nfactors = 5, rotate = "varimax")

# Print the orthogonal rotation factor loadings
print(train_FA_orth$loadings, cutoff = 0.001, digits = 3, sort = TRUE)
```

```
Loadings:
    MR1   MR3   MR5   MR2   MR4
Q6D 0.733 0.098 0.057 0.053 0.100
Q6E 0.612 0.180 0.151 0.077 0.243
Q6F 0.684 0.147 0.062 0.046 0.152
Q6N 0.656 0.048 0.108 0.045 0.281
Q6G 0.164 0.917 0.137 0.110 0.166
Q6H 0.210 0.826 0.099 0.119 0.149
Q6K 0.129 0.072 0.702 0.214 0.124
Q6M 0.090 0.084 0.773 0.126 0.104
Q6J 0.066 0.093 0.233 0.961 0.123
Q6B 0.199 0.067 0.102 0.096 0.741
Q6C 0.243 0.115 0.127 0.098 0.656
Q6A 0.219 0.219 0.082 0.047 0.387
Q6I 0.193 0.107 0.412 0.417 0.123
Q6L 0.024 0.156 0.415 0.467 0.086

                 MR1   MR3   MR5   MR2   MR4
SS loadings    2.094 1.712 1.590 1.437 1.413
Proportion Var 0.150 0.122 0.114 0.103 0.101
Cumulative Var 0.150 0.272 0.385 0.488 0.589
```

```
# Display the plot
fa.diagram(train_FA_orth,
           main = "5 factor orthogonal (varimax)")
```

5 factor orthogonal (varimax)



Five-Factor Model with Varimax Rotation

MR1 shows high loadings for the importance of clear signage and information for passengers. MR2 has moderate loadings with attributes related to the parking facilities and shuttle services, along with roadway signs. MR3 indicates that this factor is closely associated with the availability of directional aids and information booths. MR4 captures dining and shopping options, and MR5 encompasses rental cars and AirTrain services. Together, these five factors explain 59% of the total variance in passenger satisfaction, suggesting that these identified factors can explain a substantial amount of the variance in passenger satisfaction.

Next, we explored the Four-Factor EFA Model with Promax Rotation.

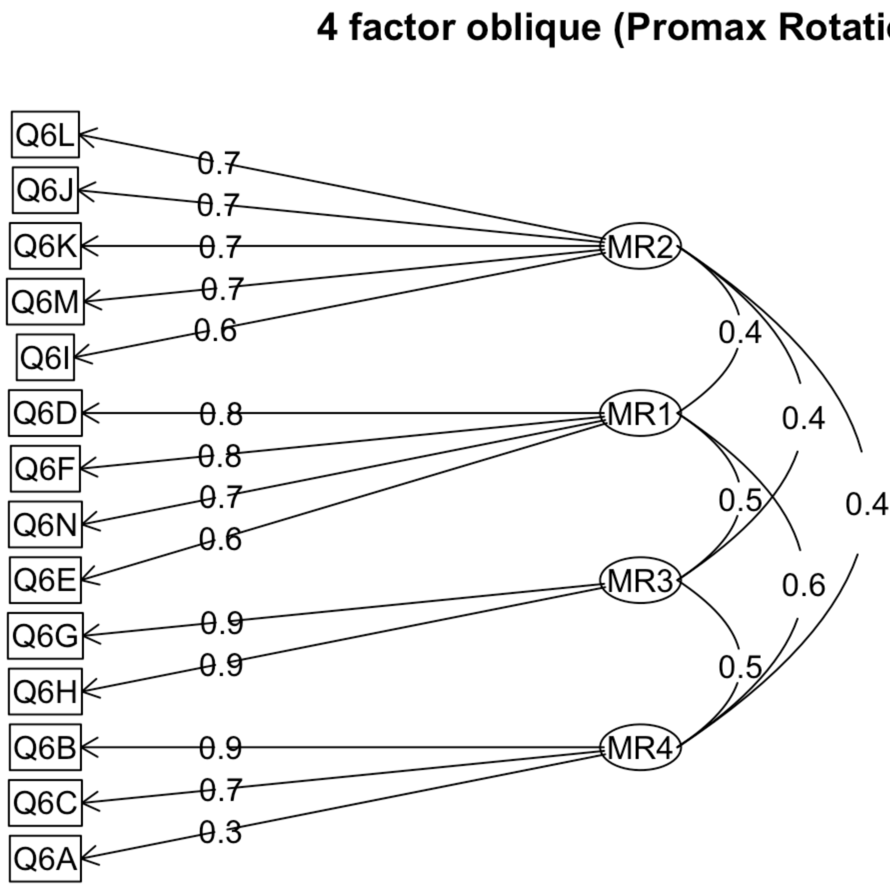
```
############ Clean data for Promax ############
customer_satisfaction_data_clean <- dat %>% dplyr::select(Q6A:Q6N) %>% drop_na()
set.seed(1842)
row_indices_clean <- 1:nrow(customer_satisfaction_data_clean)
train_indices_clean <- sample(row_indices_clean, size = 0.7 * length(row_indices_clean))
train_data_clean <- customer_satisfaction_data_clean[train_indices_clean, ]
test_data_clean <- customer_satisfaction_data_clean[-train_indices_clean, ]
nrow(train_data_clean) # Number of rows in training data
```

```
[1] 1932
```

```r
############## four-factor EFA model Promax ##############

# Fit the five-factor FA model with oblique rotation (promax):
train_data_clean_FA_obliq_4 <- fa(train_data_clean, nfactors = 4, rotate = "promax") #typ

# Print the oblique rotation factor loadings
print(train_data_clean_FA_obliq_4$loadings, cutoff = 0.001, digits = 3, sort = TRUE)
```

```
Loadings:
    MR2    MR1    MR3    MR4
Q6I  0.604  0.107
Q6J  0.712 -0.104  0.037  0.053
Q6K  0.701  0.089 -0.053 -0.042
Q6L  0.723 -0.124  0.058 -0.014
Q6M  0.653  0.031 -0.036  0.003
Q6D -0.043  0.817 -0.005 -0.115
Q6E  0.059  0.617  0.060  0.050
Q6F -0.010  0.769  0.021 -0.059
Q6N -0.004  0.657 -0.091  0.158
Q6G  0.032 -0.042  0.941 -0.017
Q6H -0.021  0.030  0.941 -0.067
Q6B -0.014 -0.090 -0.100  0.915
Q6C  0.025  0.038 -0.022  0.684
Q6A -0.017  0.095  0.165  0.345

                 MR2   MR1   MR3   MR4
SS loadings    2.321 2.140 1.828 1.478
Proportion Var 0.166 0.153 0.131 0.106
Cumulative Var 0.166 0.319 0.449 0.555
```

```r
# Display the plot
fa.diagram(train_data_clean_FA_obliq_4,
           main = "4 factor oblique (Promax Rotation)")
```



Four-Factor Model with Promax Rotation

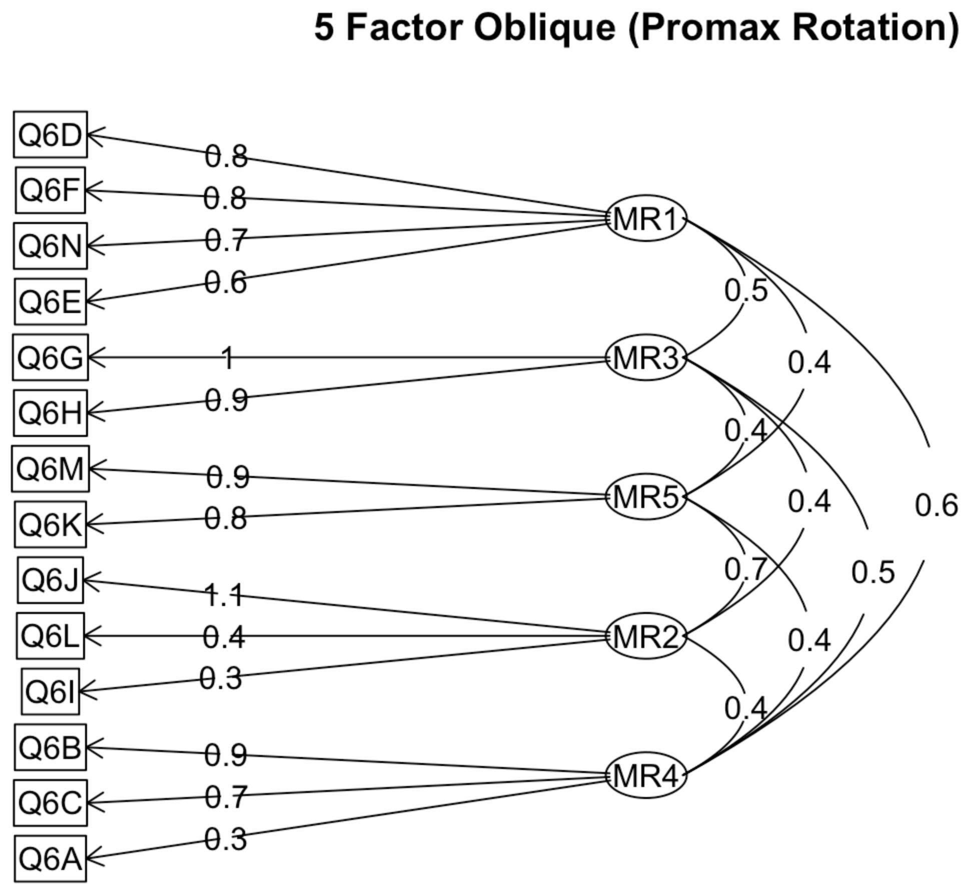
MR1 indicates that passengers highly value clear signage and screen information. MR2 represents various transportation services available at the airport, such as parking facilities, AirTrain, and rental cars, implying that the convenience and availability of transportation significantly affect passenger experiences. MR3 loads heavily on this factor, demonstrating that passengers appreciate effective assistance and the presence of information booths. MR4 shows that food and shopping options available in the airport are desirable. Together, these four factors explain a cumulative variance of 55.5%, suggesting that they cover a significant portion of the underlying structure of passenger satisfaction.

Lastly, we examined the Five-Factor EFA Model with Promax Rotation.

```
############## five-factor EFA model Promax ##############

# Fit the five-factor FA model with oblique rotation (promax):
train_data_clean_FA_obliq <- fa(train_data_clean, nfactors = 5, rotate = "promax") #typic

# Print the oblique rotation factor loadings
print(train_data_clean_FA_obliq$loadings, cutoff = 0.001, digits = 3, sort = TRUE)
```

```
Loadings:
    MR1    MR3    MR5    MR2    MR4   
Q6D  0.829 -0.013 -0.043 -0.011 -0.120
Q6E  0.618  0.060  0.052         0.050
Q6F  0.778  0.015 -0.013 -0.007 -0.062
Q6N  0.661 -0.095  0.006 -0.014  0.157
Q6G -0.069  0.996  0.027 -0.033 -0.018
Q6H  0.035  0.917 -0.046  0.004 -0.060
Q6K  0.038 -0.027  0.762 -0.010 -0.026
Q6M -0.059 -0.008  0.929 -0.178  0.020
Q6J -0.024 -0.048 -0.181  1.135  0.013
Q6B -0.098 -0.098 -0.002 -0.007  0.920
Q6C  0.039 -0.021  0.004  0.026  0.681
Q6A  0.092  0.168 -0.012 -0.011  0.345
Q6I  0.123  0.003  0.300  0.322  0.005
Q6L -0.091  0.057  0.313  0.424 -0.009

                 MR1   MR3   MR5   MR2   MR4
SS loadings    2.166 1.890 1.673 1.607 1.480
Proportion Var 0.155 0.135 0.119 0.115 0.106
Cumulative Var 0.155 0.290 0.409 0.524 0.630
```

```
# Display the plot
fa.diagram(train_data_clean_FA_obliq,
            main = "5 Factor Oblique (Promax Rotation)")
```



Five-Factor Model with Promax Rotation

MR1 indicates that passengers highly value clear signage and screen information. MR2 reflects the importance of transportation services with significant value on efficient parking options. MR3 shows strong loadings on information desks and personnel available for guidance. MR4 shows positive loadings, indicating that passengers value the quality and variety of food and shopping options available. MR5 reflects the availability and quality of transportation options like rental cars and the AirTrain, which are crucial for passenger satisfaction. Together, these five factors account for 63% of the total variance in passenger satisfaction, indicating that these factors effectively capture the underlying structure of the data.

Correlation Matrices for Factor Models

```
# Correlation matrix for 5-factor model
print(train_data_clean_FA_obliq$Phi)
```

```
          MR1       MR3       MR5       MR2       MR4
MR1 1.0000000 0.4882815 0.3755352 0.2903103 0.6215602
MR3 0.4882815 1.0000000 0.3793947 0.3761907 0.4791967
MR5 0.3755352 0.3793947 1.0000000 0.6879922 0.4028087
MR2 0.2903103 0.3761907 0.6879922 1.0000000 0.3564025
MR4 0.6215602 0.4791967 0.4028087 0.3564025 1.0000000
```

```r
# Correlation matrix for 4-factor model
print(train_data_clean_FA_obliq_4$Phi)
```

```
          MR2       MR1       MR3       MR4
MR2 1.0000000 0.3617198 0.3903541 0.4273307
MR1 0.3617198 1.0000000 0.4696729 0.6144230
MR3 0.3903541 0.4696729 1.0000000 0.4733325
MR4 0.4273307 0.6144230 0.4733325 1.0000000
```

Correlation matrices from both Promax Rotation models provide additional insights into factor interrelationships. They reveal how different aspects of customer experience are linked, allowing for targeted interventions that can improve overall satisfaction.

In the Five-Factor model, MR1 (Navigating Inside) has a strong correlation of 0.6216 with MR4 (Navigating Outside), suggesting that enhancements to signage could lead to increased overall satisfaction. The correlation of 0.6880 between MR2 (Parking and Shuttles) and MR5 (Airport and Rental Services) indicates that improving parking facilities may enhance transportation services. In the Four-Factor model, the correlation of 0.6144 between Overall Satisfaction and Amenities remains significant, underscoring the interdependencies between these factors.

Confirmatory Factor Analysis (CFA) on Passenger Satisfaction

We conducted a Confirmatory Factor Analysis (CFA) on the test dataset, using the best models identified from our Exploratory Factor Analysis (EFA). This analysis focuses on the Five-Factor Model with Varimax Rotation and the Five-Factor EFA Model with Promax Rotation. The purpose of this step is to confirm whether the factors identified in our EFA accurately represent the underlying constructs and relationships among the observed variables. CFA tests the hypothesis that the relationships between the observed variables and their latent factors hold true in a separate sample. By assessing the model's fit, we can ensure its robustness, thereby providing greater confidence in the generalizability of our findings.

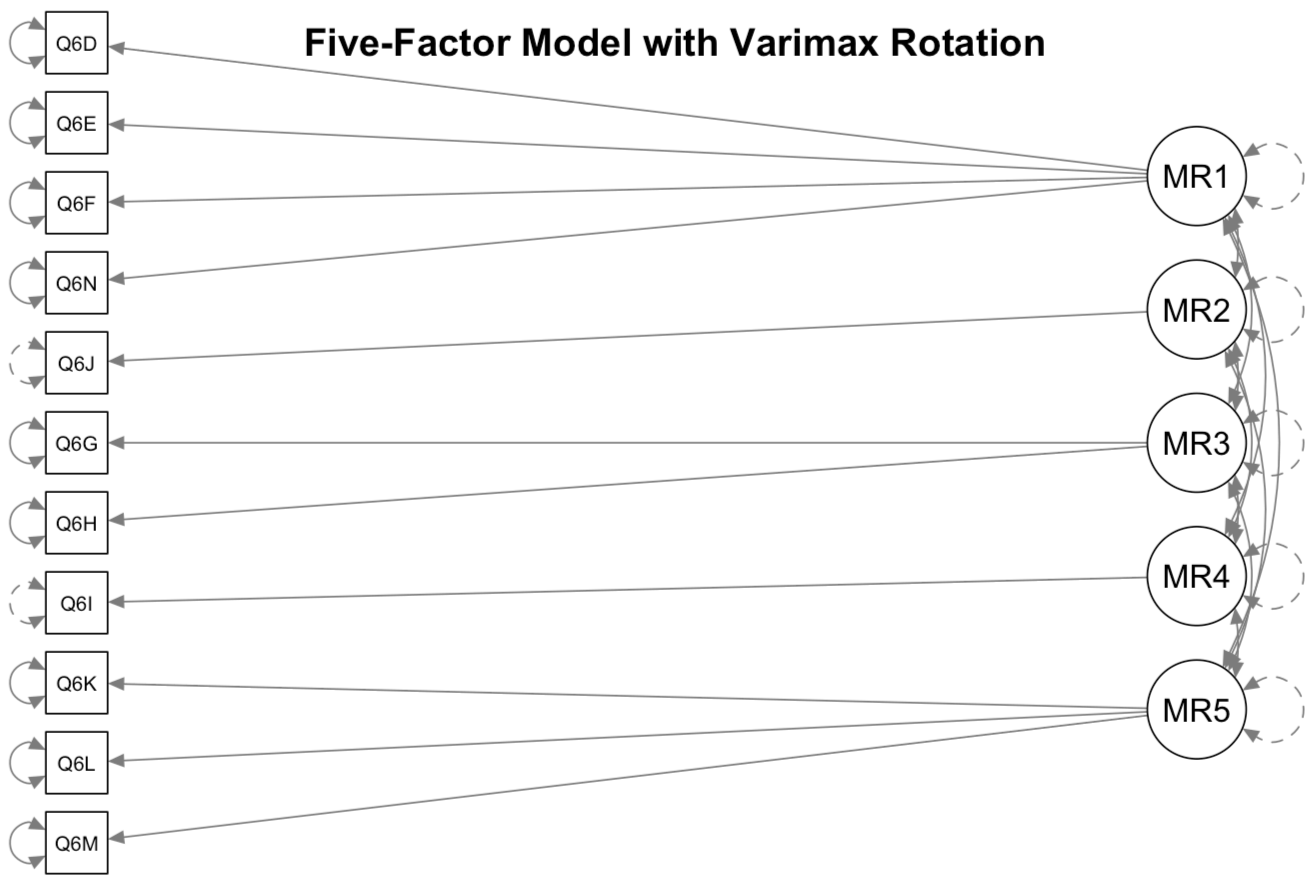
```r
# Define the CFA model for Five-Factor Model with Varimax Rotation
model_five_varimax <- "
  MR1 =~ Q6D + Q6E + Q6F + Q6N
  MR2 =~ Q6J
  MR3 =~ Q6G + Q6H
  MR4 =~ Q6I
  MR5 =~ Q6K + Q6L + Q6M
"

# Fit the CFA model
fit_five_varimax <- lavaan::cfa(model_five_varimax, data = test_data, std.lv = TRUE)

# Model Summary
summary(fit_five_varimax, fit.measures = T, standardized = T)
```

```
lavaan 0.6-19 ended normally after 28 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        30

                                                  Used       Total
  Number of observations                           855         971

Model Test User Model:

  Test statistic                               225.774
  Degrees of freedom                                36
  P-value (Chi-square)                           0.000

Model Test Baseline Model:

  Test statistic                              3831.454
  Degrees of freedom                                55
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.950
  Tucker-Lewis Index (TLI)                       0.923

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -12075.730
  Loglikelihood unrestricted model (H1)     -11962.843

  Akaike (AIC)                               24211.461
  Bayesian (BIC)                             24353.994
  Sample-size adjusted Bayesian (SABIC)      24258.722

Root Mean Square Error of Approximation:

  RMSEA                                          0.079
  90 Percent confidence interval - lower         0.069
  90 Percent confidence interval - upper         0.088
  P-value H_0: RMSEA <= 0.050                    0.000
  P-value H_0: RMSEA >= 0.080                    0.415

Standardized Root Mean Square Residual:

  SRMR                                           0.040

Parameter Estimates:

  Standard errors                             Standard
```

```
  Information                                 Expected
  Information saturated (h1) model          Structured

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  MR1 =~
    Q6D               0.659    0.030   21.987    0.000    0.659    0.718
    Q6E               0.689    0.031   22.423    0.000    0.689    0.729
    Q6F               0.711    0.031   23.276    0.000    0.711    0.750
    Q6N               0.479    0.025   19.434    0.000    0.479    0.651
  MR2 =~
    Q6J               1.161    0.028   41.352    0.000    1.161    1.000
  MR3 =~
    Q6G               1.184    0.040   29.576    0.000    1.184    0.923
    Q6H               1.167    0.040   29.397    0.000    1.167    0.918
  MR4 =~
    Q6I               1.211    0.029   41.352    0.000    1.211    1.000
  MR5 =~
    Q6K               0.738    0.037   20.000    0.000    0.738    0.681
    Q6L               0.712    0.035   20.444    0.000    0.712    0.694
    Q6M               0.779    0.042   18.705    0.000    0.779    0.644

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  MR1 ~~
    MR2               0.180    0.037    4.883    0.000    0.180    0.180
    MR3               0.448    0.033   13.559    0.000    0.448    0.448
    MR4               0.374    0.033   11.234    0.000    0.374    0.374
    MR5               0.302    0.042    7.174    0.000    0.302    0.302
  MR2 ~~
    MR3               0.298    0.033    9.125    0.000    0.298    0.298
    MR4               0.515    0.025   20.469    0.000    0.515    0.515
    MR5               0.682    0.026   26.742    0.000    0.682    0.682
  MR3 ~~
    MR4               0.295    0.033    9.009    0.000    0.295    0.295
    MR5               0.378    0.038   10.052    0.000    0.378    0.378
  MR4 ~~
    MR5               0.607    0.029   21.247    0.000    0.607    0.607

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .Q6D               0.409    0.026   15.682    0.000    0.409    0.485
   .Q6E               0.418    0.027   15.341    0.000    0.418    0.469
   .Q6F               0.392    0.027   14.602    0.000    0.392    0.437
   .Q6N               0.312    0.018   17.259    0.000    0.312    0.576
   .Q6J               0.000                               0.000    0.000
   .Q6G               0.244    0.054    4.502    0.000    0.244    0.148
   .Q6H               0.253    0.053    4.795    0.000    0.253    0.157
   .Q6I               0.000                               0.000    0.000
   .Q6K               0.630    0.040   15.767    0.000    0.630    0.536
   .Q6L               0.547    0.036   15.392    0.000    0.547    0.519
```

```
.Q6M              0.857    0.051   16.693    0.000    0.857    0.585
 MR1              1.000                               1.000    1.000
 MR2              1.000                               1.000    1.000
 MR3              1.000                               1.000    1.000
 MR4              1.000                               1.000    1.000
 MR5              1.000                               1.000    1.000
```

```r
# Visualizing the Model
semPaths(fit_five_varimax, title = FALSE, rotation = 4, mar = c(2, 2, 2, 2))
title("Five-Factor Model with Varimax Rotation")
```




```r
fit.index = c("chisq", "df", "pvalue", "cfi", "tli", "rmsea", "srmr","gfi") # Name Fit Me

# Show condensed Fit
fitMeasures(fit_five_varimax, fit.index)
```

```
  chisq      df  pvalue     cfi     tli   rmsea    srmr     gfi 
225.774  36.000   0.000   0.950   0.923   0.079   0.040   0.951 
```

Five-Factor Model with Varimax Rotation

The CFA for the Five-Factor Model yielded a chi-square statistic of 225.774 with 36 degrees of freedom, resulting in a significant p-value of 0.000. This suggests that while the model does not perfectly fit the data, it is essential to note that the chi-square statistic is highly sensitive to sample size. Therefore, we considered additional fit indices for a more nuanced evaluation.

The Comparative Fit Index (CFI) was 0.950, indicating a good fit, as values above 0.95 generally indicate a strong model fit. The Tucker-Lewis Index (TLI) was 0.923, suggesting an acceptable fit but falling slightly short of the ideal threshold of 0.95. The Root Mean Square Error of Approximation (RMSEA) was 0.079, with an upper confidence interval reaching 0.088, indicating an acceptable but not excellent fit. Lastly, the Goodness-of-Fit Index (GFI) was 0.951, further supporting the model's reasonably good representation of the underlying factor structure. Overall, the strong values for CFI, SRMR, and GFI demonstrate that this model effectively captures the factors in the data. However, the RMSEA and TLI suggest minor room for improvement.

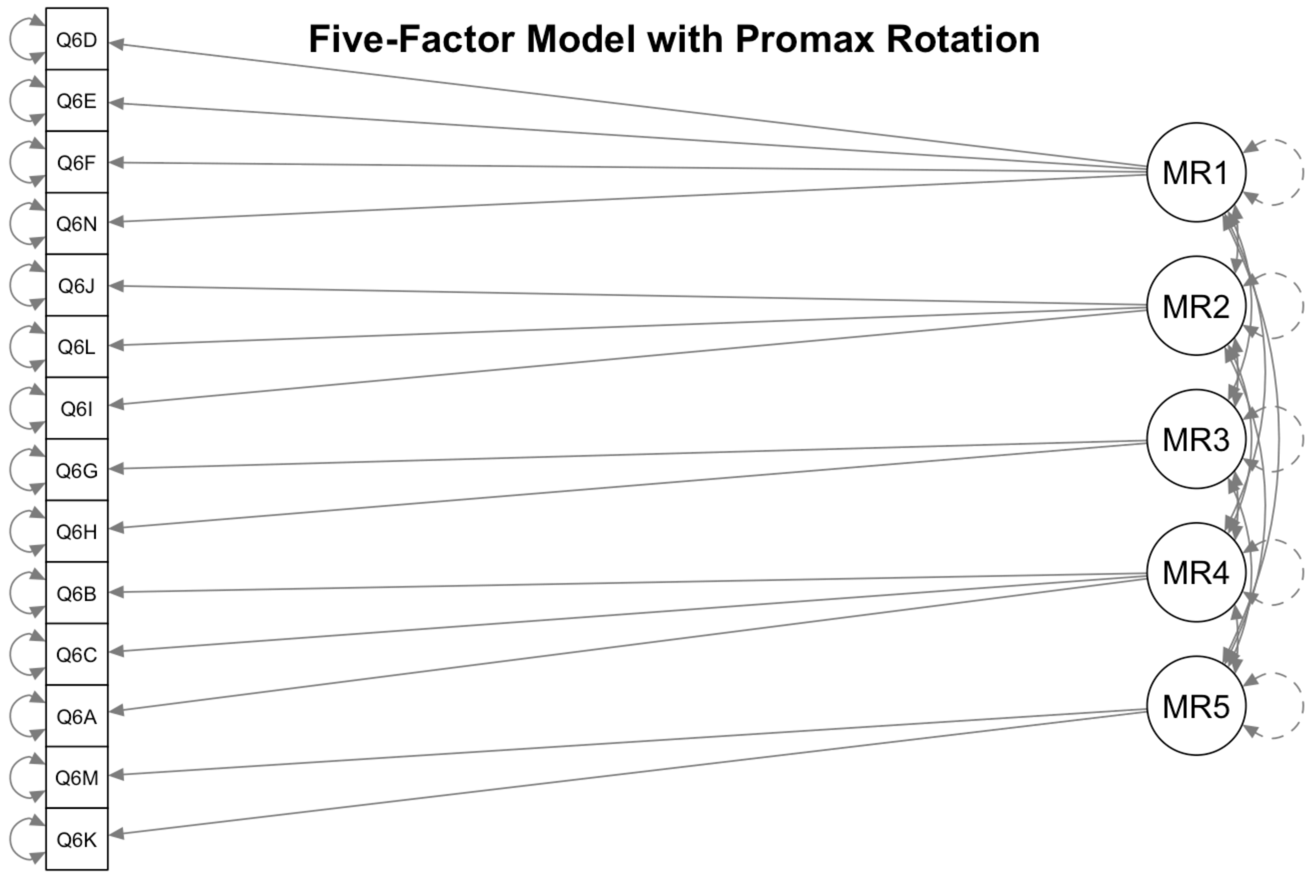
```r
############# Five-Factor Model with Promax Rotation #############

# Define the CFA model for Five-Factor Model with Promax Rotation
model_five_promax <- "
  MR1 =~ Q6D + Q6E + Q6F + Q6N
  MR2 =~ Q6J + Q6L + Q6I
  MR3 =~ Q6G + Q6H
  MR4 =~ Q6B + Q6C + Q6A
  MR5 =~ Q6M + Q6K
"

# Fit the CFA model
fit_five_promax <- lavaan::cfa(model_five_promax, data = test_data, std.lv = TRUE)

# Model Summary
summary(fit_five_promax, fit.measures = T, standardized = T)
```

```
lavaan 0.6-19 ended normally after 27 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        38

                                                  Used       Total
  Number of observations                           831         971

Model Test User Model:

  Test statistic                               301.971
  Degrees of freedom                                67
  P-value (Chi-square)                           0.000

Model Test Baseline Model:

  Test statistic                              4640.154
  Degrees of freedom                                91
  P-value                                        0.000
```

```
User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.948
  Tucker-Lewis Index (TLI)                       0.930

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -15447.627
  Loglikelihood unrestricted model (H1)     -15296.642

  Akaike (AIC)                               30971.255
  Bayesian (BIC)                             31150.715
  Sample-size adjusted Bayesian (SABIC)      31030.040

Root Mean Square Error of Approximation:

  RMSEA                                          0.065
  90 Percent confidence interval - lower         0.058
  90 Percent confidence interval - upper         0.073
  P-value H_0: RMSEA <= 0.050                    0.000
  P-value H_0: RMSEA >= 0.080                    0.000

Standardized Root Mean Square Residual:

  SRMR                                           0.049

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  MR1 =~
    Q6D               0.664    0.030   22.088    0.000    0.664    0.724
    Q6E               0.692    0.031   22.291    0.000    0.692    0.729
    Q6F               0.700    0.031   22.672    0.000    0.700    0.738
    Q6N               0.495    0.025   20.067    0.000    0.495    0.672
  MR2 =~
    Q6J               0.916    0.038   23.917    0.000    0.916    0.787
    Q6L               0.745    0.035   21.427    0.000    0.745    0.719
    Q6I               0.781    0.042   18.794    0.000    0.781    0.646
  MR3 =~
    Q6G               1.220    0.039   31.043    0.000    1.220    0.948
    Q6H               1.141    0.039   28.907    0.000    1.141    0.896
  MR4 =~
    Q6B               0.868    0.044   19.679    0.000    0.868    0.694
    Q6C               0.932    0.045   20.860    0.000    0.932    0.732
    Q6A               0.680    0.045   14.971    0.000    0.680    0.547
```

```
  MR5 =~
    Q6M               0.859    0.044   19.493    0.000    0.859    0.708
    Q6K               0.828    0.040   20.821    0.000    0.828    0.761

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  MR1 ~~
    MR2               0.303    0.041    7.426    0.000    0.303    0.303
    MR3               0.436    0.034   12.963    0.000    0.436    0.436
    MR4               0.661    0.032   20.787    0.000    0.661    0.661
    MR5               0.332    0.042    7.813    0.000    0.332    0.332
  MR2 ~~
    MR3               0.408    0.035   11.511    0.000    0.408    0.408
    MR4               0.457    0.040   11.400    0.000    0.457    0.457
    MR5               0.747    0.031   24.355    0.000    0.747    0.747
  MR3 ~~
    MR4               0.504    0.034   14.615    0.000    0.504    0.504
    MR5               0.323    0.040    8.160    0.000    0.323    0.323
  MR4 ~~
    MR5               0.400    0.044    9.074    0.000    0.400    0.400

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .Q6D               0.402    0.026   15.587    0.000    0.402    0.476
   .Q6E               0.422    0.027   15.439    0.000    0.422    0.469
   .Q6F               0.409    0.027   15.150    0.000    0.409    0.455
   .Q6N               0.298    0.018   16.824    0.000    0.298    0.549
   .Q6J               0.516    0.042   12.207    0.000    0.516    0.381
   .Q6L               0.518    0.035   14.960    0.000    0.518    0.483
   .Q6I               0.853    0.051   16.840    0.000    0.853    0.583
   .Q6G               0.168    0.052    3.215    0.001    0.168    0.101
   .Q6H               0.319    0.048    6.678    0.000    0.319    0.197
   .Q6B               0.810    0.056   14.467    0.000    0.810    0.518
   .Q6C               0.754    0.058   13.050    0.000    0.754    0.465
   .Q6A               1.080    0.061   17.754    0.000    1.080    0.700
   .Q6M               0.735    0.056   13.177    0.000    0.735    0.499
   .Q6K               0.497    0.046   10.712    0.000    0.497    0.421
    MR1               1.000                               1.000    1.000
    MR2               1.000                               1.000    1.000
    MR3               1.000                               1.000    1.000
    MR4               1.000                               1.000    1.000
    MR5               1.000                               1.000    1.000
```

```r
# Visualizing the Model
semPaths(fit_five_promax,, title = FALSE, rotation = 4, mar = c(2, 2, 2, 2))
title("Five-Factor Model with Promax Rotation")
```




```
# Show condensed Fit
fitMeasures(fit_five_promax, fit.index)
```

```
  chisq      df  pvalue     cfi     tli   rmsea    srmr     gfi
301.971  67.000   0.000   0.948   0.930   0.065   0.049   0.951
```

Five-Factor Model with Promax Rotation

For the Five-Factor Model with Promax Rotation, we found a chi-square statistic of 301.971 with 67 degrees of freedom, yielding a significant p-value of 0.000. The CFI was 0.948, indicating a good fit, while the TLI was 0.923, suggesting an acceptable fit. The RMSEA was consistent at 0.065, implying an acceptable fit. The GFI for this model was 0.951, reinforcing that this model adequately represents the underlying factor structure as well. Similar to the Five-Factor Model with Varimax Rotation, these results indicate that while the model is robust, minor room remains for enhancement.

Model Modifications

We further examined our models by obtaining modification indices for both five-factor models, which helped identify potential improvements.

```
# Get modification indices for the Five-Factor Model with Varimax Rotation
modificationIndices(fit_five_varimax,  minimum.value = 10, sort = TRUE)
```

	lhs	op	rhs	mi	epc	sepc.lv	sepc.all	sepc.nox
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
53	MR2	=~	Q6L	156.21370	0.7158544	0.7158544	0.6970921	0.6970921
120	Q6J	~~	Q6L	146.91304	0.3878577	0.3878577	NA	NA
135	Q6K	~~	Q6M	64.50404	0.3001767	0.3001767	0.4086358	0.4086358
121	Q6J	~~	Q6M	54.18845	-0.2700357	-0.2700357	NA	NA
54	MR2	=~	Q6M	53.91560	-0.4759575	-0.4759575	-0.3934484	-0.3934484
52	MR2	=~	Q6K	29.61328	-0.3251990	-0.3251990	-0.3000426	-0.3000426
132	Q6I	~~	Q6L	28.69859	-0.1705815	-0.1705815	NA	NA
119	Q6J	~~	Q6K	26.53227	-0.1725193	-0.1725193	NA	NA
134	Q6K	~~	Q6L	21.02646	-0.1545090	-0.1545090	-0.2631910	-0.2631910
72	MR4	=~	Q6L	19.85780	-0.2108455	-0.2108455	-0.2053194	-0.2053194

1-10 of 14 rows Previous 1 2 Next

```
# Get modification indices for the Five-Factor Model with Promax Rotation
modificationIndices(fit_five_promax, minimum.value = 10, sort = TRUE)
```

	lhs	op	r...	mi	epc	sepc.lv	sepc.all	sepc.nox
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
155	Q6L	~~	Q6I	62.00349	-0.27582774	-0.27582774	-0.4148622	-0.4148622
46	MR1	=~	Q6I	59.31023	0.33230132	0.33230132	0.2747728	0.2747728
92	MR5	=~	Q6J	56.42920	-0.71361016	-0.71361016	-0.6127640	-0.6127640
146	Q6J	~~	Q6L	50.99825	0.29920565	0.29920565	0.5783036	0.5783036
94	MR5	=~	Q6I	42.24679	0.57490805	0.57490805	0.4753790	0.4753790
83	MR4	=~	Q6I	33.39515	0.28990433	0.28990433	0.2397156	0.2397156
153	Q6J	~~	Q6M	21.22151	-0.15644814	-0.15644814	-0.2539138	-0.2539138
74	MR3	=~	Q6A	20.80756	0.23720585	0.23720585	0.1910676	0.1910676
80	MR4	=~	Q6N	20.01221	0.17623598	0.17623598	0.2391438	0.2391438
181	Q6B	~~	Q6C	18.92929	0.27856003	0.27856003	0.3563103	0.3563103

1-10 of 22 rows Previous 1 2 3 Next

- Varimax Rotation: The modification indices suggested adding a covariance between Parking Facilities (Q6J) and Shuttle to Long-Term Parking (Q6L).
- Promax Rotation: The modification indices indicated the need to add covariances between Shuttle Parking (Q6L) and Roadway Signs (Q6I), as well as between Parking (Q6J) and Shuttle Parking (Q6L).

```
# Update the CFA model for Five-Factor Model with Varimax Rotation
mod_model_five_varimax <- "
  MR1 =~ Q6D + Q6E + Q6F + Q6N
  MR2 =~ Q6J + Q6L
  MR3 =~ Q6G + Q6H
  MR4 =~ Q6I
```

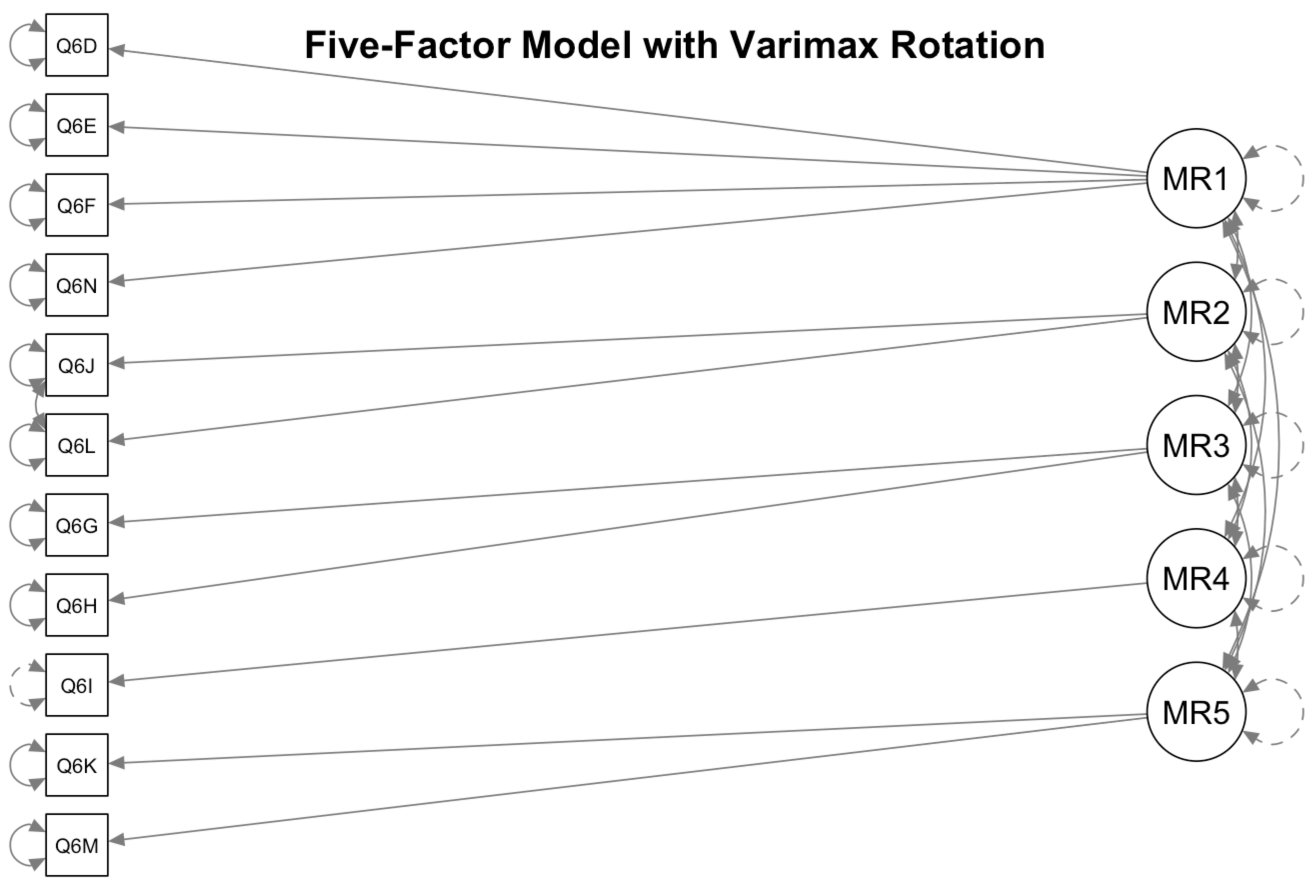
```r
  MR5 =~ Q6K + Q6M
  Q6J ~~ Q6L
"

# Fit the CFA model
fit_five_varimax_mod <- lavaan::cfa(mod_model_five_varimax, data = test_data, std.lv = TR

# Model Summary
summary(fit_five_varimax_mod, fit.measures = T, standardized = T)
```

```
lavaan 0.6-19 ended normally after 29 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        32

                                                  Used       Total
  Number of observations                           855         971

Model Test User Model:

  Test statistic                               113.837
  Degrees of freedom                                34
  P-value (Chi-square)                           0.000

Model Test Baseline Model:

  Test statistic                              3831.454
  Degrees of freedom                                55
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.979
  Tucker-Lewis Index (TLI)                       0.966

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -12019.762
  Loglikelihood unrestricted model (H1)     -11962.843

  Akaike (AIC)                               24103.524
  Bayesian (BIC)                             24255.559
  Sample-size adjusted Bayesian (SABIC)      24153.936

Root Mean Square Error of Approximation:

  RMSEA                                          0.052
  90 Percent confidence interval - lower         0.042
  90 Percent confidence interval - upper         0.063
```

```
  P-value H_0: RMSEA <= 0.050                    0.337
  P-value H_0: RMSEA >= 0.080                    0.000

Standardized Root Mean Square Residual:

  SRMR                                           0.027

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  MR1 =~                                                                
    Q6D               0.659       NA                      0.659    0.717
    Q6E               0.688       NA                      0.688    0.728
    Q6F               0.711       NA                      0.711    0.751
    Q6N               0.479       NA                      0.479    0.651
  MR2 =~                                                                
    Q6J               0.927       NA                      0.927    0.798
    Q6L               0.726       NA                      0.726    0.706
  MR3 =~                                                                
    Q6G               1.186       NA                      1.186    0.925
    Q6H               1.164       NA                      1.164    0.917
  MR4 =~                                                                
    Q6I               1.211       NA                      1.211    1.000
  MR5 =~                                                                
    Q6K               0.830       NA                      0.830    0.766
    Q6M               0.856       NA                      0.856    0.708

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
 .Q6J ~~                                                                
   .Q6L               0.054       NA                      0.054    0.106
  MR1 ~~                                                                
    MR2               0.216       NA                      0.216    0.216
    MR3               0.447       NA                      0.447    0.447
    MR4               0.374       NA                      0.374    0.374
    MR5               0.326       NA                      0.326    0.326
  MR2 ~~                                                                
    MR3               0.381       NA                      0.381    0.381
    MR4               0.599       NA                      0.599    0.599
    MR5               0.693       NA                      0.693    0.693
  MR3 ~~                                                                
    MR4               0.295       NA                      0.295    0.295
    MR5               0.326       NA                      0.326    0.326
  MR4 ~~                                                                
    MR5               0.589       NA                      0.589    0.589
```

```
Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .Q6D               0.411       NA                      0.411    0.486
   .Q6E               0.419       NA                      0.419    0.470
   .Q6F               0.390       NA                      0.390    0.435
   .Q6N               0.312       NA                      0.312    0.576
   .Q6J               0.489       NA                      0.489    0.363
   .Q6L               0.528       NA                      0.528    0.501
   .Q6G               0.238       NA                      0.238    0.145
   .Q6H               0.258       NA                      0.258    0.160
   .Q6I               0.000                               0.000    0.000
   .Q6K               0.486       NA                      0.486    0.414
   .Q6M               0.730       NA                      0.730    0.499
    MR1               1.000                               1.000    1.000
    MR2               1.000                               1.000    1.000
    MR3               1.000                               1.000    1.000
    MR4               1.000                               1.000    1.000
    MR5               1.000                               1.000    1.000
```

```r
# Visualizing the Model
semPaths(fit_five_varimax_mod, title = FALSE, rotation = 4, mar = c(2, 2, 2, 2))
title("Five-Factor Model with Varimax Rotation")
```



```r
fit.index = c("chisq", "df", "pvalue", "cfi", "tli", "rmsea", "srmr","gfi") # Name Fit Me

# Show condensed Fit
fitMeasures(fit_five_varimax_mod, fit.index)
```

```
  chisq      df  pvalue     cfi     tli   rmsea    srmr     gfi 
113.837  34.000   0.000   0.979   0.966   0.052   0.027   0.976 
```

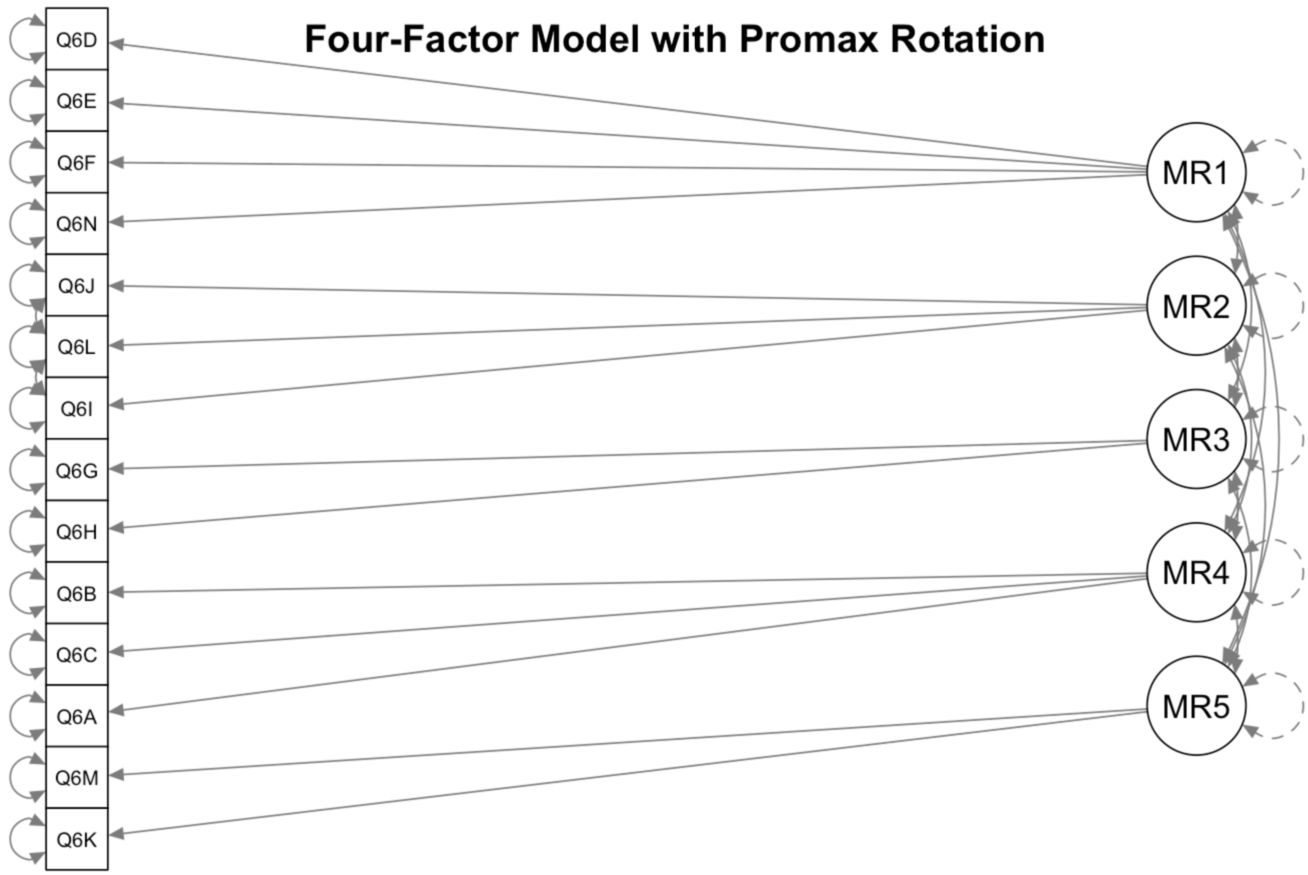
```r
############# Five-Factor Model with Promax Rotation #############

# Update the CFA model for Five-Factor Model with Promax Rotation
mod_model_five_promax <- "
  MR1 =~ Q6D + Q6E + Q6F + Q6N
  MR2 =~ Q6J + Q6L + Q6I
  MR3 =~ Q6G + Q6H
  MR4 =~ Q6B + Q6C + Q6A
  MR5 =~ Q6M + Q6K
  Q6L ~~ Q6I
  Q6J ~~ Q6L
"

# Fit the CFA model
fit_five_promax_mod <- lavaan::cfa(mod_model_five_promax, data = test_data, std.lv = TRUE

# Model Summary
summary(fit_five_promax_mod, fit.measures = T, standardized = T)
```

```
lavaan 0.6-19 ended normally after 30 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        40

                                                  Used       Total
  Number of observations                           831         971

Model Test User Model:
                                                      
  Test statistic                               213.772
  Degrees of freedom                                65
  P-value (Chi-square)                           0.000

Model Test Baseline Model:

  Test statistic                              4640.154
  Degrees of freedom                                91
  P-value                                        0.000

User Model versus Baseline Model:
```

```
  Comparative Fit Index (CFI)                    0.967
  Tucker-Lewis Index (TLI)                       0.954

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)             -15403.528
  Loglikelihood unrestricted model (H1)     -15296.642

  Akaike (AIC)                               30887.056
  Bayesian (BIC)                             31075.961
  Sample-size adjusted Bayesian (SABIC)      30948.935

Root Mean Square Error of Approximation:

  RMSEA                                          0.052
  90 Percent confidence interval - lower         0.045
  90 Percent confidence interval - upper         0.060
  P-value H_0: RMSEA <= 0.050                    0.289
  P-value H_0: RMSEA >= 0.080                    0.000

Standardized Root Mean Square Residual:

  SRMR                                           0.040

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  MR1 =~
    Q6D               0.664    0.030   22.080    0.000    0.664    0.723
    Q6E               0.693    0.031   22.340    0.000    0.693    0.730
    Q6F               0.700    0.031   22.659    0.000    0.700    0.738
    Q6N               0.495    0.025   20.049    0.000    0.495    0.671
  MR2 =~
    Q6J               0.775    0.042   18.457    0.000    0.775    0.665
    Q6L               0.720    0.045   15.900    0.000    0.720    0.695
    Q6I               0.938    0.044   21.240    0.000    0.938    0.775
  MR3 =~
    Q6G               1.221    0.039   30.945    0.000    1.221    0.949
    Q6H               1.140    0.040   28.773    0.000    1.140    0.896
  MR4 =~
    Q6B               0.869    0.044   19.691    0.000    0.869    0.695
    Q6C               0.932    0.045   20.827    0.000    0.932    0.731
    Q6A               0.680    0.045   14.974    0.000    0.680    0.548
  MR5 =~
    Q6M               0.874    0.043   20.224    0.000    0.874    0.720
```

```
    Q6K               0.813    0.039   20.967    0.000    0.813    0.748

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
 .Q6L ~~
   .Q6I              -0.213    0.040   -5.382    0.000   -0.213   -0.375
 .Q6J ~~
   .Q6L               0.180    0.045    3.995    0.000    0.180    0.277
  MR1 ~~
    MR2               0.371    0.039    9.459    0.000    0.371    0.371
    MR3               0.436    0.034   12.951    0.000    0.436    0.436
    MR4               0.661    0.032   20.790    0.000    0.661    0.661
    MR5               0.331    0.043    7.777    0.000    0.331    0.331
  MR2 ~~
    MR3               0.406    0.036   11.328    0.000    0.406    0.406
    MR4               0.475    0.040   11.884    0.000    0.475    0.475
    MR5               0.782    0.033   23.785    0.000    0.782    0.782
  MR3 ~~
    MR4               0.504    0.034   14.612    0.000    0.504    0.504
    MR5               0.325    0.040    8.208    0.000    0.325    0.325
  MR4 ~~
    MR5               0.404    0.044    9.177    0.000    0.404    0.404

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .Q6D               0.402    0.026   15.602    0.000    0.402    0.477
   .Q6E               0.421    0.027   15.413    0.000    0.421    0.467
   .Q6F               0.409    0.027   15.170    0.000    0.409    0.456
   .Q6N               0.299    0.018   16.839    0.000    0.299    0.550
   .Q6J               0.756    0.051   14.969    0.000    0.756    0.558
   .Q6L               0.554    0.056    9.966    0.000    0.554    0.517
   .Q6I               0.583    0.058   10.081    0.000    0.583    0.399
   .Q6G               0.165    0.053    3.120    0.002    0.165    0.100
   .Q6H               0.321    0.048    6.642    0.000    0.321    0.198
   .Q6B               0.809    0.056   14.434    0.000    0.809    0.517
   .Q6C               0.756    0.058   13.068    0.000    0.756    0.466
   .Q6A               1.079    0.061   17.746    0.000    1.079    0.700
   .Q6M               0.709    0.054   13.155    0.000    0.709    0.481
   .Q6K               0.521    0.044   11.868    0.000    0.521    0.440
    MR1               1.000                               1.000    1.000
    MR2               1.000                               1.000    1.000
    MR3               1.000                               1.000    1.000
    MR4               1.000                               1.000    1.000
    MR5               1.000                               1.000    1.000
```

```r
# Visualizing the Model
semPaths(fit_five_promax_mod, labels = factor_descriptions, title = FALSE, rotation = 4,
title("Four-Factor Model with Promax Rotation")
```




```
# Show condensed Fit
fitMeasures(fit_five_promax_mod, fit.index)
```

```
  chisq      df  pvalue     cfi     tli   rmsea    srmr     gfi 
213.772  65.000   0.000   0.967   0.954   0.052   0.040   0.964 
```

Updated Models

Five-Factor Model with Varimax Rotation

After updating the Five-Factor Model with Varimax Rotation to include the covariance (Q6J ~~ Q6L), we fitted the updated model, resulting in significantly improved fit indices. The chi-square statistic was 113.837 with 34 degrees of freedom, yielding a significant p-value of 0.000 due to the large sample size. The comparative fit index (CFI) was 0.979, indicating a very good fit, while the Tucker-Lewis index (TLI) was 0.966, suggesting a good fit. Additionally, the standardized root mean square residual (SRMR) was 0.027, and the goodness-of-fit index (GFI) was 0.976, supporting the model's strong fit to the data.

Five-Factor Model with Promax Rotation

After incorporating the covariances (Q6L ~~ Q6I and Q6J ~~ Q6L) into the Five-Factor Model with Promax Rotation, we fitted the updated model, which showed significantly improved fit indices. The chi-square statistic for this model was 213.772 with 65 degrees of freedom, again yielding a significant p-value of 0.000. The CFI was 0.967, indicating a very good fit, and the TLI was 0.954, suggesting a good fit. The root mean square error of approximation (RMSEA) was 0.052, indicating an acceptable fit. The GFI for this model was 0.964, further supporting its strong fit to the data.

Summary

Overall, the updated CFA models for the Five-Factor Model with Varimax Rotation and the Five-Factor Model with Promax Rotation demonstrated improved fit indices, suggesting that the modifications made were appropriate and enhanced the models' representation of the underlying data. Notably, the Five-Factor Model with Varimax Rotation exhibited better fit characteristics, indicating that this structure more effectively captures the underlying data patterns and provides a more robust understanding of the latent constructs at play.

Factor Scores from Confirmatory Factor Analysis

After successfully completing our Confirmatory Factor Analysis (CFA), we obtained factor scores for the Five-Factor Model with Varimax Rotation. Factor scores serve as quantitative representations of the underlying latent variables for each observation in our dataset. Essentially, these scores indicate how each participant scores on the identified factors derived from the CFA, providing valuable insights into the relationships among observed variables.

Using factor scores enables us to explore further assumptions related to demographic variables included in the survey, such as age, gender, and household income. By examining these scores in relation to demographic characteristics, we can uncover patterns and relationships that may inform our understanding of passenger satisfaction. This section outlines three specific analyses conducted using the factor scores: a t-test comparing factor scores across income groups, a one-way ANOVA analyzing the differences in factor scores based on gender, and a two-way ANOVA investigating the interaction effects of gender and age group on the factor scores. These analyses will help clarify how demographic factors influence perceptions of service quality and facilities, enhancing our findings.

```
# Obtain factor scores from the Five-Factor Model with Varimax Rotation
predFacScores <- data.frame(lavPredict(fit_five_varimax_mod))

# Preview the factor scores
round(head(predFacScores),3)
```

	MR1	MR2	MR3	MR4	MR5
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.580	0.348	0.908	-0.567	-0.185
2	0.102	0.866	0.897	1.085	0.833
3	0.676	0.739	0.918	1.085	0.465
4	-0.781	-1.768	-1.495	-0.567	-0.801
5	0.447	0.847	0.912	1.085	0.849
6	0.691	-0.155	0.903	-2.218	-1.245

6 rows

We utilized the lavPredict() function from the lavaan package in R to compute the factor scores for the updated model, reflecting the modifications made based on the modification indices. The resulting factor scores provide insights into how each observation relates to the identified factors. For example, the first observation has relatively high scores on MR1 (Navigating Inside) and MR3 (information Booths) but a negative score on MR4 (Navigating Outside) and a slightly negative score on MR5 (Airport and Rental Services). This suggests that this observation is strongly associated with the factors represented by MR1 and MR3 but less so with MR4 and MR5.

In contrast, the fourth observation has negative scores on all factors, especially MR2 (Parking and Shuttles) and MR3, indicating a weaker association with these identified factors and potentially representing an outlier or a very different pattern in terms of how it loads on these factors. Overall, the factor scores reflect how each observation aligns with the factors derived from the revised model, capturing more accurate associations based on the improvements suggested by the modification indices.

Integrating Demographic Data with Factor Scores

Our next analysis will integrate customer demographic data with the factor scores generated from our updated factor model. The goal is to leverage both sets of information—demographics and factor scores—to understand data patterns better.

We selected specific demographic columns from the dataset for further analysis: age (Q17), gender (Q18), and household income (Q19). These demographic variables help us understand how different groups of customers relate to the identified factors from our Confirmatory Factor Analysis (CFA). To verify our selection, we displayed the first few rows of the chosen columns, ensuring that we captured the relevant demographic data accurately.

```
# Select specific demographic columns
customer_demographics <- dat[, c("Q17", "Q18", "Q19")]

# Display the rows
head(customer_demographics)
```

	Q17	Q18	Q19
	<int>	<int>	<int>
1	4	1	1
2	2	2	1
3	7	2	2
4	5	2	3
5	6	2	1
6	6	2	2

6 rows

```
# Set a seed for reproducibility
set.seed(1842)

# Create a vector of row indices and randomly sample 70% for the training set
train_indices <- sample(1:nrow(customer_demographics), size = 0.7 * nrow(customer_demogra

# Create the training and test sets
demo_train_data <- customer_demographics[train_indices, ]
demo_test_data <- customer_demographics[-train_indices, ]

# Verify the split
# head(demo_train_data)
# head(demo_test_data)
# nrow(demo_train_data)  # Number of rows in training data
# nrow(demo_test_data)   # Number of rows in test data

# Merge test set with factor scores
dat_with_scores <- cbind(demo_test_data[1:nrow(predFacScores), ], predFacScores)

# Verify the merge
head(dat_with_scores)
```

	Q...	Q...	Q...	MR1	MR2	MR3	MR4	MR5
	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2	2	2	1	0.5797743	0.3481579	0.9081753	-0.5669006	-0.1851293
3	7	2	2	0.1021049	0.8657370	0.8974997	1.0845475	0.8330707
6	6	2	2	0.6764255	0.7387493	0.9180934	1.0845475	0.4646652
7	2	1	1	-0.7810094	-1.7684346	-1.4951018	-0.5669006	-0.8009445
14	7	2	4	0.4473886	0.8466118	0.9119887	1.0845475	0.8489836
15	4	2	4	0.6908730	-0.1550816	0.9032489	-2.2183487	-1.2449694

6 rows

Next, we set a seed value to guarantee the reproducibility of our random sampling process. We created a vector of row indices and randomly sampled 70% of the dataset for the training set, reserving the remaining 30% for the test set. This split resulted in 2,263 rows in the training set and 971 rows in the test set, which we confirmed.

This structured dataset enables us to explore the relationships between customer demographics and their respective factor scores. Analyzing these connections could reveal valuable insights into how different demographic groups perceive and evaluate various aspects of airport services, ultimately enhancing our understanding of passenger satisfaction.

Income Groups

We started by analyzing the relationship between income levels and the factor scores obtained from our Confirmatory Factor Analysis (CFA). Specifically, we wanted to determine if there are significant differences in the factor scores between high (income over $50,000) and low (income below $50,000) income groups. To accomplish this, we formulated the following hypotheses:

- H0: There is no significant difference in the factor scores between high and low-income groups.
- H1: There is a significant difference in the factor scores between high and low-income groups.

Mathematically, we can express these hypotheses as:

- $H_0 : \mu_{low} = \mu_{high}$ vs. $H_1 : \mu_{low} \neq \mu_{high}$

To facilitate our analysis, we collapsed the income variable into categories: high, low, and other. We then conducted a Welch Two-Sample t-test to compare the factor scores between the high and low-income groups for each factor.

```
# Collapse income into categories (high, low. other)
dat_with_scores$income_cat <- factor(ifelse(dat_with_scores$Q19 == 1, "Low", ifelse(dat_w

# Perform t-test for the first factor
t_test_MR1 <- t.test(dat_with_scores$MR1 ~ dat_with_scores$income_cat)

# View the t-test result
t_test_MR1
```

```
	Welch Two Sample t-test

data:  dat_with_scores$MR1 by dat_with_scores$income_cat
t = -0.87544, df = 330, p-value = 0.382
alternative hypothesis: true difference in means between group Low and group High is not
equal to 0
95 percent confidence interval:
 -0.22280622  0.08557114
sample estimates:
 mean in group Low mean in group High 
       -0.04676404         0.02185350 
```

The t-test results for the first factor, MR1, yielded a p-value of 0.382. This value is greater than our significance level of 0.05, leading us to fail to reject the null hypothesis. This outcome indicates that there is insufficient evidence to support a significant difference in MR1 scores between the high and low-income groups.

We extended this analysis to factors MR2 through MR5, performing t-tests for each. Similar to MR1, the results revealed no statistically significant differences between the two income groups across these factors. Thus, we can interpret these findings to suggest that whether a respondent belongs to the high or low-income group, their satisfaction and perceptions related to navigation (inside or outside), parking and shuttles, information booths, and AirTrain and rental services are statistically similar.

Age Groups

Next, we examined the relationship between age groups and the factor scores. Specifically, we wanted to determine if there are significant differences in the factor scores across different age groups. To achieve this, we formulated the following hypotheses:

- H0: There are no significant differences in factor scores across different age groups.
- H1: There are significant differences in factor scores across different age groups.

Mathematically, we can express these hypotheses as

- $H_0 : \mu_{\text{Under 18}} = \mu_{18\text{–}24} = \mu_{25\text{–}34} = \mu_{35\text{–}44} = \mu_{45\text{–}54} = \mu_{55\text{–}64} = \mu_{\text{65 and over}} = \mu_{\text{Don't know/Refused}}$
  vs.
  $H_1 : \mu_{\text{Under 18}} \neq \mu_{18\text{–}24} \neq \mu_{25\text{–}34} \neq \mu_{35\text{–}44} \neq \mu_{45\text{–}54} \neq \mu_{55\text{–}64} \neq \mu_{\text{65 and over}} \neq \mu_{\text{Don't know/Refused}}$

We first created a new variable representing the different age groups to conduct this analysis. Subsequently, we performed a one-way ANOVA to compare the factor scores across these age groups for the first factor (MR1).

```r
# Create a new variable for age groups
dat_with_scores$age_group <- factor(dat_with_scores$Q17, levels = c(1, 2, 3, 4, 5, 6, 7,

# Remove rows with missing values from the dataset
# dat_with_scores <- na.omit(dat_with_scores)

# Perform one-way ANOVA for the first factor
anova_MR1 <- aov(MR1 ~ age_group, data = dat_with_scores)


# View the ANOVA result
summary(anova_MR1)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
age_group     6    6.9  1.1475   1.392  0.215
Residuals   825  679.9  0.8242
23 observations deleted due to missingness
```

The results of the one-way ANOVA for MR1 yielded a p-value of 0.215, which exceeds our significance level of 0.05. Consequently, we fail to reject the null hypothesis, indicating that age groups do not have a significant effect on MR1 scores.

We extended this analysis to factors MR2 through MR5, conducting additional one-way ANOVA tests. The results consistently revealed no statistically significant differences among the age groups across these factors. This further supports the conclusion that age does not significantly influence factor scores. Thus, we can interpret these findings to mean that a respondent's age does not significantly affect their perceptions of navigation (inside or outside), parking and shuttles, information booths, and AirTrain and rental services.

Age and Gender Interaction

In our final analysis, we wanted to analyze the relationship between age groups, gender, and their interaction in relation to the factor scores obtained. Specifically, we wanted to determine whether there are significant differences in the factor scores based on age group, gender, or their interaction. To facilitate this analysis, we formulated the following hypotheses:

- H0: There are no significant differences in factor scores based on age group, gender, or their interaction (all means are equal).
- H1: There are significant differences in factor scores based on age group, gender, or their interaction (at least one mean is different).

Mathematically, we can express these hypotheses as

- Age:
  $H_0 : \mu_{\text{Under 18}} = \mu_{18\text{–}24} = \mu_{25\text{–}34} = \mu_{35\text{–}44} = \mu_{45\text{–}54} = \mu_{55\text{–}64} = \mu_{\text{65 and over}} = \mu_{\text{Don't know/Refused}}$
  vs,
  $H_1 : \mu_{\text{Under 18}} \neq \mu_{18\text{–}24} \neq \mu_{25\text{–}34} \neq \mu_{35\text{–}44} \neq \mu_{45\text{–}54} \neq \mu_{55\text{–}64} \neq \mu_{\text{65 and over}} \neq \mu_{\text{Don't know/Refused}}$
- Gender: $H_0 : \mu_{low} = \mu_{high}$ vs. $H_1 : \mu_{low} \neq \mu_{high}$
- Interaction:$H_{0,\text{interaction}}$ vs $H_{1,\text{interaction}}$

We began by creating a new variable categorizing respondents into gender groups and performed a two-way ANOVA to compare the factor scores based on age group, gender, and their interaction for the first factor (MR1).

```
# Create a new variable for gender groups
dat_with_scores$gender <- factor(dat_with_scores$Q18, levels = c(1, 2), labels = c("Male"

# Remove any rows with NA values
dat_with_scores_filtered <- na.omit(dat_with_scores[, c("MR1", "MR2", "MR3", "MR4", "MR5"

# Perform Two-Way ANOVA
anova_MR1 <- aov(MR1 ~ age_group * gender, data = dat_with_scores_filtered)

# View the summary of the ANOVA
summary(anova_MR1)
```

```
                  Df Sum Sq Mean Sq F value Pr(>F)
age_group          6    7.6  1.2701   1.527  0.166
gender             1    0.1  0.1374   0.165  0.685
age_group:gender   6    3.7  0.6126   0.737  0.620
Residuals        788  655.3  0.8316
```

The results of the two-way ANOVA test showed the p-values for the factors in MR1 were 0.166 for the age group, 0.685 for gender, and 0.620 for the interaction between age and gender, all of which are greater than the significance level (alpha = 0.05). Therefore, we fail to reject the null hypothesis, indicating that these factors do not significantly affect MR1 scores.

We also conducted additional two-way ANOVA tests for MR2 through MR5. These tests similarly revealed no statistically significant differences across age, gender, or their interaction. This supports our conclusion that age, gender, and their interaction do not appear to significantly influence navigation (inside or outside), parking and shuttles, information booths, and AirTrain and rental services.

Sentiment Analysis

In exploring customer feedback regarding airport services, we conducted a comprehensive sentiment analysis to uncover insights into traveler sentiments. We constructed a new data frame incorporating key columns such as respondent identifiers, quantitative satisfaction ratings, qualitative text responses, and relevant demographic information.

The selected columns are:

- RESPNum: Respondent number, allowing us to track individual responses.
- Q6A to Q6N: Questions related to various airport features, capturing quantitative ratings of customer satisfaction.
- SAQ: Administration of the survey, providing context about how the data was collected.
- Q7_text_All: Predetermined text responses that offer qualitative feedback, critical for understanding sentiment nuances.
- Demographic Information (Q17, Q18, Q19): Age group, gender, and income level.

```
# Creating new df with columns needed for sentiment analysis
dat2 <- dat[, c("RESPNUM", "Q6A", "Q6B", "Q6C", "Q6D", "Q6E", "Q6F",
  "Q6G", "Q6H", "Q6I", "Q6J", "Q6K", "Q6L", "Q6M", "Q6N", "SAQ", "Q7_text_All",
  "Q17", "Q18", "Q19")]
head(dat2)
```

	RESPNUM <chr>	Q6A <int>	Q6B <int>	Q6C <int>	Q6D <int>	Q6E <int>	Q6F <int>	Q6G <int>	Q6H <int>
1	1	5	3	3	4	4	4	6	6
2	2	6	3	3	4	4	6	6	6
3	3	6	6	3	4	3	5	6	6
4	4	3	6	6	3	4	3	6	6
5	5	5	5	5	5	4	5	5	5
6	6	6	6	6	4	5	5	6	6

6 rows | 1-10 of 21 columns

To perform sentiment analysis on customer feedback, we first examined the structure of the relevant text data in the Q7_text_All column, which contains responses from survey participants. Using the str function, we confirmed that this column consists of character data with 3,234 entries.

To ensure the accuracy of our sentiment analysis, we took several steps to clean the text data. We identified and addressed any empty strings within the Q7_text_All column by replacing them with NA values. Additionally, we trimmed any leading or trailing whitespace to maintain consistency. Finally, we filtered out rows containing NA values in the Q7_text_All column, ensuring our analysis was based solely on text responses.

```
# Check the structure of the dataset
str(dat2$Q7_text_All)
```

```
 chr [1:3234] " Airport Layout-Terminals too far away/confusing/too difficult/takes too
long to get around airport/no way to g"| __truncated__ ...
```

```
# Check dimensions
# dim(dat$Q7_text_All)

head(dat2$Q7_text_All, 10)
```

```
 [1] " Airport Layout-Terminals too far away/confusing/too difficult/takes too long to
get around airport/no way to get between terminals after getting through
security/domestic airlines in international terminal  Smoking areas-more
needed/inconveniently located  Add traffic delays/other information to flysfo.com to
check before leaving"
 [2] " Food too expensive  Positive comment about wifi  Security/Customs lines/procedures
long/inefficient/ineffective"
 [3] " Signage outside airport confusing/hard to get to airport/difficult to find correct
terminal"
 [4] " Security/Customs lines/procedures long/inefficient/ineffective  Signage inside
airport confusing/small/hard to find gate or airline"
 [5] " Security/Customs lines/procedures long/inefficient/ineffective"
 [6] ""
 [7] " General cleanliness neg comment  Need more artwork/exhibitions/change artwork more
frequently"
 [8] ""
 [9] ""
[10] " Wifi-not free long enough/difficult to access/doesn’t cover entire
airport/overloaded/didn’t know was available"
```

```
# Replace empty strings in Q7_text_All with NA
dat2 <- dat2 %>%
  mutate(Q7_text_All = ifelse(Q7_text_All == "", NA, Q7_text_All),
         Q7_text_All = trimws(Q7_text_All))

# Filter out NA for accuracy
dat2 <- dat2 %>%
  filter(!is.na(Q7_text_All))
```

Sentiment Distribution

After preparing the dataset, we computed the average sentiment score for each text response using the sentiment_by function. This function analyzes the sentiment of each sentence within the text. It calculates an overall sentiment score for each response, providing a quantitative measure of sentiment.

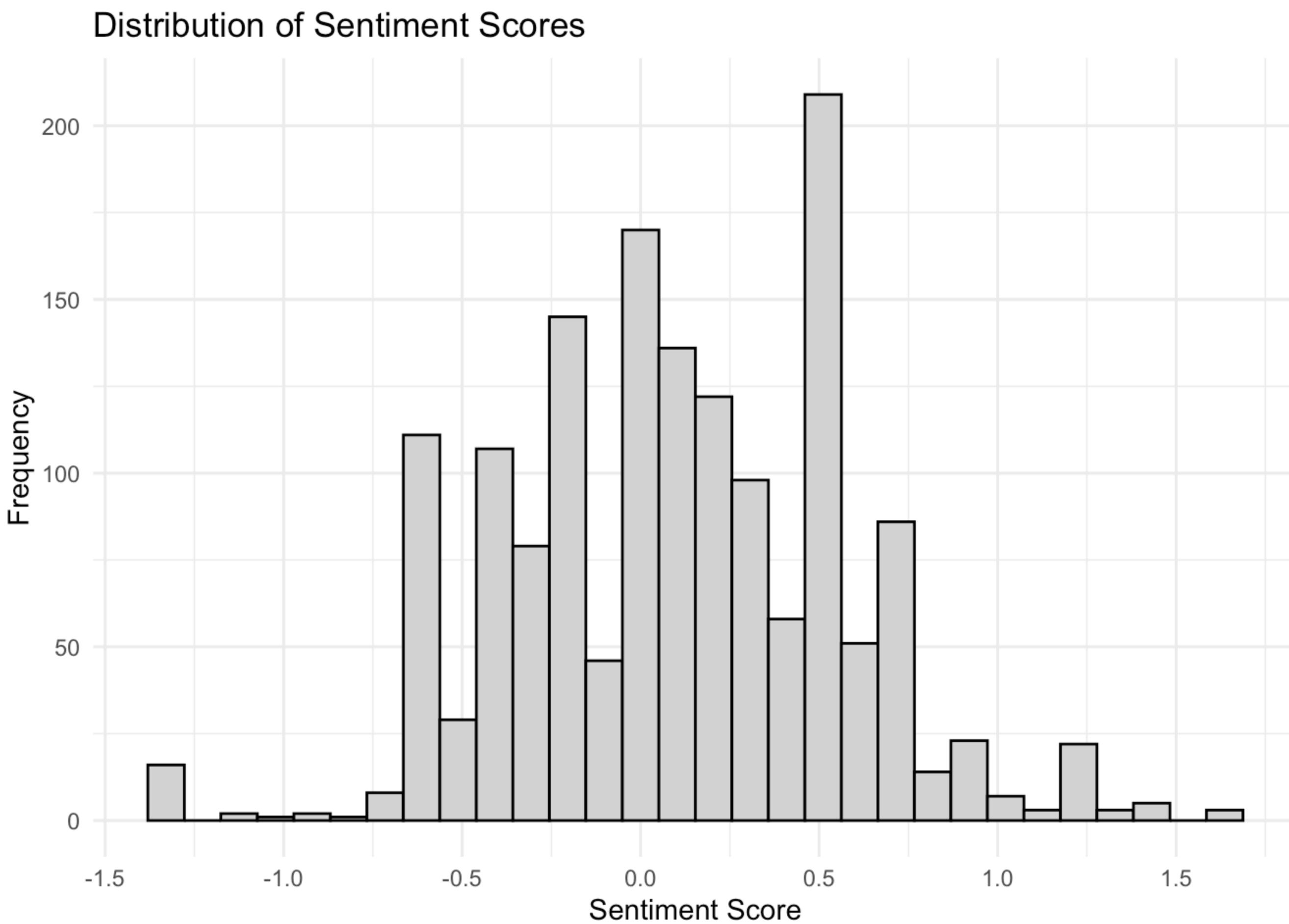
Next, we visualized the distribution of sentiment scores using the ggplot2 package. This plot allows us to see the spread of sentiment across all responses. Additionally, we generated a summary of the sentiment scores, which helps us better understand the distribution and central tendencies, such as the mean, median, and any potential outliers.

```
# Compute single (average) sentiment score without NA's
single_sentiment_score <- sentiment_by(get_sentences(dat2$Q7_text_All))
round(single_sentiment_score, 5)
```

	element_id <dbl>	word_count <dbl>	sd <dbl>	ave_sentiment <dbl>
1	1	49	0.46265	-0.35703
2	2	14	NA	-0.16036
3	3	14	NA	-0.24054
4	4	18	NA	-0.53033
5	5	7	NA	-0.56695
6	6	12	NA	1.02191
7	7	19	NA	0.18353
8	8	11	NA	-0.22613
9	9	44	NA	-0.47488
10	10	7	NA	-0.56695

1-10 of 1,557 rows Previous 1 2 3 4 5 6 ... 156 Next

```
# Plot the distribution of sentiment scores
ggplot(single_sentiment_score, aes(x = ave_sentiment)) +
geom_histogram(fill = "lightgrey", color = "black") +
  scale_x_continuous(breaks = seq(-1.5, 1.7, by = 0.5)) +
  labs(title = "Distribution of Sentiment Scores",
       x ="Sentiment Score",
       y ="Frequency") +
  theme_minimal()
```




```
summary(single_sentiment_score)
```

```
  element_id     word_count          sd          ave_sentiment
 Min.   :   1   Min.   : 3.00   Min.   :0.2006   Min.   :-1.28072
 1st Qu.: 390   1st Qu.: 7.00   1st Qu.:0.2661   1st Qu.:-0.22613
 Median : 779   Median :11.00   Median :0.3316   Median : 0.07071
 Mean   : 779   Mean   :13.99   Mean   :0.3316   Mean   : 0.10090
 3rd Qu.:1168   3rd Qu.:19.00   3rd Qu.:0.3971   3rd Qu.: 0.51430
 Max.   :1557   Max.   :63.00   Max.   :0.4626   Max.   : 1.68464
                                NA's   :1555
```

We observed that the distribution of sentiment scores appears slightly right-skewed, indicating a prevalence of more positive sentiment scores. The center of the distribution is around 0.05, suggesting that most reviews are neutral or slightly positive. The median sentiment score is 0.07, while the mean is 0.10, further reinforcing the slight positive skew in the data. The overall distribution ranges from -1.28 to 1.68, with most sentiment scores falling between -0.05 and 0.05. This indicates that while most reviews are close to neutral, there is a tendency toward slight positivity overall.

Survey Administration

We aimed to determine whether the method of survey administration could influence the sentiment expressed in customer feedback. Specifically, we examined three modes of survey delivery: interviewer-administered (1), self-administered (2), and unknown administration method (0). The way these surveys are administered might affect sentiment scores, as the presence of an interviewer or the lack of direct interaction in a self-administered survey could influence how respondents express their opinions, potentially introducing bias or shaping the overall tone of their responses.

```
# Combine data into a new data frame
combined_data <- data.frame(
     element_id = single_sentiment_score$element_id,
     word_count = single_sentiment_score$word_count,
     sd = NA,
     ave_sentiment = round(single_sentiment_score$ave_sentiment,5),
     admin = dat2$SAQ
   )
combined_data
```

element_id	word_count	sd	ave_sentiment	admin
<int>	<int>	<lgl>	<dbl>	<int>
1	49	NA	-0.35703	1
2	14	NA	-0.16036	1
3	14	NA	-0.24054	1
4	18	NA	-0.53033	1
5	7	NA	-0.56695	1
6	12	NA	1.02191	2
7	19	NA	0.18353	2
8	11	NA	-0.22613	2
9	44	NA	-0.47488	1
10	7	NA	-0.56695	1

1-10 of 1,557 rows Previous **1** 2 3 4 5 6 ... 156 Next

```
summary_by_survey <- combined_data %>%
     group_by(admin) %>%
     summarise(
       mean_sentiment = round(mean(ave_sentiment, na.rm = TRUE),4),
       median_sentiment = round(median(ave_sentiment, na.rm = TRUE),4)
     )
summary_by_survey
```

admin	mean_sentiment	median_sentiment
<int>	<dbl>	<dbl>
1	0.0938	0.0632
2	0.1057	0.0866

2 rows

We combined the sentiment scores with the survey administration type into a new dataframe and summarized the sentiment scores by administration type to observe any differences. For interviewer-administered surveys, the mean sentiment score was 0.0938, with a median of 0.0632. In contrast, self-administered surveys showed a slightly higher mean sentiment score of 0.1057 and a median of 0.0866. Both groups exhibited marginally positive sentiment on average, with self-administered responses being slightly more positive. The minimal differences between the mean and median sentiment scores across these groups suggest that the method of survey administration does not significantly impact the sentiment of the comments.

To further investigate the relationship between survey administration method and sentiment, we created binary variables for both administration type and sentiment.

```
# Create binary variables for SAQ and sentiment
# 1 for interviewer-administered and 0 for self-administered.
combined_data$admin_binary <- ifelse(combined_data$admin == 1, 1,
  ifelse(combined_data$admin == 2, 0, NA))

# 1 for positive sentiment and 0 for negative sentiment.
combined_data$sentiment_binary <- ifelse(combined_data$ave_sentiment > 0, 1, 0)

# Generate and print the confusion matrix
confusion_matrix <- confusionMatrix(
     factor(combined_data$sentiment_binary),
     factor(combined_data$admin_binary)
   )
confusion_matrix
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 408 284
         1 519 346

               Accuracy : 0.4843
                 95% CI : (0.4592, 0.5094)
    No Information Rate : 0.5954
    P-Value [Acc > NIR] : 1

                  Kappa : -0.0101

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.4401
            Specificity : 0.5492
         Pos Pred Value : 0.5896
         Neg Pred Value : 0.4000
             Prevalence : 0.5954
         Detection Rate : 0.2620
   Detection Prevalence : 0.4444
      Balanced Accuracy : 0.4947
```

```
'Positive' Class : 0
```

```
# Calculate the correlation between average sentiment and recommendation
cor(combined_data$sentiment_binary, combined_data$admin_binary, use = "complete.obs")
```

```
[1] -0.01053361
```

We then printed the confusion matrix to evaluate the model's performance. The matrix revealed a model accuracy of 0.4843, indicating low performance, no better than random guessing. Additionally, the correlation coefficient of -0.0105 suggested an almost negligible relationship between survey administration method and sentiment.

Income and Sentiment Scores

Our next objective was to evaluate if there are significant differences in the mean sentiment scores between the "Low" and "High" income groups based on sentiment. To achieve this, we formulated the following hypotheses

- H0: There is no significant difference in the mean sentiment scores between the "Low" and "High" income groups.
- H1: There is a significant difference in the mean sentiment scores between the "Low" and "High" income groups.

Mathematically, we can express these hypotheses as

- $H_0 : \mu_{low} = \mu_{high}$ vs. $H_1 : \mu_{low} \neq \mu_{high}$

```
# Collapse income into two categories (high and low)
dat2$income_cat <- factor(ifelse(dat2$Q19 == 1, "Low", ifelse(dat2$Q19 %in% c(2, 3, 4), "

# Filter the dataset to responses that match element IDs in single_sentiment_score
dat_filtered <- dat2 %>% filter(RESPNUM %in% single_sentiment_score$element_id)

# Combine the filtered data with sentiment scores
combined_data <- cbind(dat_filtered[1:nrow(single_sentiment_score), ], single_sentiment_s

# Combine data into a new data frame using RESPNUM as the identifier
combined_data <- data.frame(
  element_id = combined_data$RESPNUM,
  word_count = combined_data$word_count,
  ave_sentiment = round(combined_data$ave_sentiment, 5),
  survey_responses = combined_data$SAQ,
  income_cat = combined_data$income_cat )
combined_data
```

element_id	word_count	ave_sentiment	survey_responses	income_cat
<chr>	<int>	<dbl>	<int>	<fct>
1	49	-0.35703	1	Low
2	14	-0.16036	1	Low
3	14	-0.24054	1	High
4	18	-0.53033	1	High
5	7	-0.56695	1	Low
7	12	1.02191	2	Low
10	19	0.18353	2	High
11	11	-0.22613	2	High
15	44	-0.47488	1	High
16	7	-0.56695	1	High

1-10 of 1,557 rows Previous 1 2 3 4 5 6 ... 156 Next

```
# Perform t-test for the first factor
t_test_income <- t.test(combined_data$ave_sentiment ~ combined_data$income_cat)

# View the t-test result
t_test_income
```

```
	Welch Two Sample t-test

data:  combined_data$ave_sentiment by combined_data$income_cat
t = 1.8731, df = 131.18, p-value = 0.06328
alternative hypothesis: true difference in means between group Low and group High is not
equal to 0
95 percent confidence interval:
 -0.00618555  0.22656686
sample estimates:
 mean in group Low mean in group High
        0.16158692         0.05139627
```

We collapsed the income variable into two categories: “Low” and “High.” After filtering the dataset to include only responses matching element IDs in the single_sentiment_score dataframe, we created a new dataframe using RESPNUM as the identifier, which included the sentiment scores and corresponding income categories.

To compare the sentiment scores between the two income groups, we performed a Welch Two-Sample t-test. The test produced a p-value of 0.06328, which is greater than the significance level (alpha = 0.05). Therefore, we fail to reject the null hypothesis, indicating no significant difference in the mean sentiment scores between the “Low” and “High” income groups. This suggests that income level does not significantly influence the sentiment expressed in the comments.

Age Group and Sentiment

Following this, we aimed to examine if there are significant differences in the mean sentiment scores among different age groups. To achieve this, we formulated the following hypotheses:

- H0: There is no significant difference in the mean sentiment scores between the different age groups.
- H1: There is a significant difference in the mean sentiment scores between at least two of the age groups.

Mathematically, we can express these hypotheses as

- $H_0 : \mu_{\text{Under 18}} = \mu_{18\text{–}24} = \mu_{25\text{–}34} = \mu_{35\text{–}44} = \mu_{45\text{–}54} = \mu_{55\text{–}64} = \mu_{\text{65 and over}} = \mu_{\text{Don't know/Refused}}$
  vs.
  $H_1 : \mu_{\text{Under 18}} \neq \mu_{18\text{–}24} \neq \mu_{25\text{–}34} \neq \mu_{35\text{–}44} \neq \mu_{45\text{–}54} \neq \mu_{55\text{–}64} \neq \mu_{\text{65 and over}} \neq \mu_{\text{Don't know/Refused}}$

```
# Create a new variable for age groups based on Q17
dat2$age_group <- factor(dat2$Q17, levels = c(1, 2, 3, 4, 5, 6, 7, 8), labels = c("Under

# If the row shows NA change to not specified
dat2$age_group[is.na(dat2$age_group)] <- "Don't know/Refused"

# Filter the dataset to responses that match element IDs in single_sentiment_score
dat_filtered <- dat2 %>% filter(RESPNUM %in% single_sentiment_score$element_id)

# Combine the filtered data with sentiment scores
combined_data <- cbind(dat_filtered[1:nrow(single_sentiment_score), ], single_sentiment_s

# Combine data into a new data frame using RESPNUM as the identifier
combined_data <- data.frame(
  element_id = combined_data$RESPNUM,
  word_count = combined_data$word_count,
  ave_sentiment = round(combined_data$ave_sentiment, 5),
  survey_responses = combined_data$SAQ,
  income_cat = combined_data$income_cat,
  age_group = combined_data$age_group)
combined_data
```

element_id	word_count	ave_sentiment	survey_responses	income_cat
<chr>	<int>	<dbl>	<int>	<fct>
1	49	-0.35703	1	Low
2	14	-0.16036	1	Low
3	14	-0.24054	1	High
4	18	-0.53033	1	High
5	7	-0.56695	1	Low
7	12	1.02191	2	Low
10	19	0.18353	2	High
11	11	-0.22613	2	High
15	44	-0.47488	1	High

element_id	word_count	ave_sentiment	survey_responses	income_cat
<chr>	<int>	<dbl>	<int>	<fct>
16	7	-0.56695	1	High

1-10 of 1,557 rows | 1-5 of 6 columns Previous **1** 2 3 4 5 6 ... 156 Next

```
# Perform one-way ANOVA
anova_age <- aov(ave_sentiment ~ age_group, data = combined_data)

# View the ANOVA result
summary(anova_age)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
age_group     7   4.27  0.6106   3.127 0.00325 **
Residuals   349  68.16  0.1953
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
1200 observations deleted due to missingness
```

We created a new variable for age groups, ensuring that we handled NA values by recoding them to “Don’t know/Refused.” We then filtered the dataset to include only responses that matched the element IDs in the single_sentiment_score dataframe, which included the sentiment scores, income categories, and age groups.

To analyze the differences in sentiment scores, we performed a one-way ANOVA. The results yielded a p-value of 0.00325, which is less than the significance level (alpha = 0.05). Consequently, we reject the null hypothesis, indicating that there is a significant difference in the mean sentiment scores among the different age groups. This suggests that age group significantly influences the sentiment expressed in the comments.

Age and Gender Interaction

Lastly, we aimed to determine if there are significant differences in the mean sentiment scores based on age groups, gender, and their interaction. To achieve this, we formulated the following hypotheses:

- H0: There are no significant differences in sentiment scores based on age group, gender, or their interaction (all means are equal).
- H1: There are significant differences in sentiment scores based on age group, gender, or their interaction (at least one mean is different).

Mathematically, we can express these hypotheses as

- Age:
  $H_0 : \mu_{\text{Under 18}} = \mu_{18\text{–}24} = \mu_{25\text{–}34} = \mu_{35\text{–}44} = \mu_{45\text{–}54} = \mu_{55\text{–}64} = \mu_{\text{65 and over}} = \mu_{\text{Don't know/Refused}}$
  vs,
  $H_1 : \mu_{\text{Under 18}} \neq \mu_{18\text{–}24} \neq \mu_{25\text{–}34} \neq \mu_{35\text{–}44} \neq \mu_{45\text{–}54} \neq \mu_{55\text{–}64} \neq \mu_{\text{65 and over}} \neq \mu_{\text{Don't know/Refused}}$
- Gender: $H_0 : \mu_{low} = \mu_{high}$ vs. $H_1 : \mu_{low} \neq \mu_{high}$
- Interaction:$H_{0,\text{interaction}}$ vs $H_{1,\text{interaction}}$

We began by creating a new variable for gender groups, then we filtered the dataset to include only responses that matched element IDs in the single_sentiment_score dataframe that included sentiment scores, income categories, age groups, and gender.

```
# Create a new variable for gender groups
dat2$gender <- factor(dat2$Q18, levels = c(1, 2), labels = c("Male", "Female"))

# Filter the dataset to responses that match element IDs in single_sentiment_score
dat_filtered <- dat2 %>% filter(RESPNUM %in% single_sentiment_score$element_id)

# Combine the filtered data with sentiment scores
combined_data <- cbind(dat_filtered[1:nrow(single_sentiment_score), ], single_sentiment_s

# Combine data into a new data frame
combined_data <- data.frame(
  element_id = combined_data$RESPNUM,
  word_count = combined_data$word_count,
  ave_sentiment = round(combined_data$ave_sentiment, 5),
  survey_responses = combined_data$SAQ,
  income_cat = combined_data$income_cat,
  age_group = combined_data$age_group,
  gender = combined_data$gender)
combined_data
```

element_id <chr>	word_count <int>	ave_sentiment <dbl>	survey_responses <int>	income_cat <fct>
1	49	-0.35703	1	Low
2	14	-0.16036	1	Low
3	14	-0.24054	1	High
4	18	-0.53033	1	High
5	7	-0.56695	1	Low
7	12	1.02191	2	Low
10	19	0.18353	2	High
11	11	-0.22613	2	High
15	44	-0.47488	1	High
16	7	-0.56695	1	High

1-10 of 1,557 rows | 1-5 of 7 columns Previous 1 2 3 4 5 6 ... 156 Next

```
# Perform Two-Way ANOVA
anova_interaction <- aov(ave_sentiment ~ age_group * gender, data = combined_data)

# View the summary of the ANOVA
summary(anova_interaction)
```

```
                  Df Sum Sq Mean Sq F value  Pr(>F)
age_group          7   3.90  0.5566   2.945 0.00525 **
```

```
gender              1   1.73  1.7298   9.154 0.00268 **
age_group:gender   5   1.30  0.2599   1.375 0.23301
Residuals        332  62.74  0.1890
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
1211 observations deleted due to missingness
```

To assess the differences, we performed a two-way ANOVA. The results yielded the following p-values: for age group (0.00525) and gender (0.00268), both of which are less than the significance level (alpha = 0.05). Therefore, we reject the null hypothesis, indicating significant differences in the mean sentiment scores between different age groups and genders. However, the p-value for the interaction between age and gender (0.23301) is greater than the significance level, leading us to fail to reject the null hypothesis for the interaction. This indicates that there is no significant interaction effect between age group and gender on the mean sentiment scores.

Conclusion: Summary of Findings and Implications

The analysis conducted on passenger survey data provided valuable insights into the perceptions and sentiments of travelers regarding various airport services. Through Confirmatory Factor Analysis (CFA), we identified distinct factors that influence passenger experiences while examining the impact of demographic variables on these perceptions. Additionally, our sentiment analysis revealed nuanced patterns in how different groups express satisfaction or dissatisfaction with airport services. The following key findings synthesize the results from the CFA and sentiment analysis, highlighting the factors that significantly shape passenger experiences and uncovering potential areas for improvement in airport operations and customer service strategies.

Key Findings:

1. The Confirmatory Factor Analysis (CFA) identified five main factors:

- MR1 (Navigating Inside): Associated with: Signs and directions inside SFO, escalators, elevators, moving walkways, information on screens and monitors, and SFO Airport as a whole.
- MR2 (Parking and Shuttles): Associated with: Airport parking facilities and long-term parking lot shuttle.
- MR3 (Information Booths): Associated with: Information booths on the lower level near baggage claim and the upper level departure area.
- MR4 (Navigating Outside): Associated with: Signs and directions on SFO airport roadways.
- MR5 (AirTrain and Rental Services): Associated with: AirTrain and airport rental car center.

2. We saw insufficient evidence to suggest significant differences in MR1-MR5 factor scores between high and low-income groups. We also saw that Age groups did not significantly affect factor scores for MR1-MR5, indicating uniformity in perceptions across different age demographics, and neither gender nor the interaction between age and gender had a significant impact on MR1-MR5 scores.

3. We saw interviewer-administered and self-administered surveys exhibited marginally positive sentiment. The sentiment analysis model exhibited low accuracy (0.4843), indicating it performed no better than random guessing, suggesting the need for improved sentiment classification

methodologies. Income level did not significantly influence the sentiment expressed in comments, and there were significant differences in mean sentiment scores among different age groups, indicating that age influences how sentiment is expressed. While there were significant differences in sentiment scores between different age groups and genders, there was no significant interaction effect between age group and gender.

Implications

1. Identifying MR1 (Navigating Inside) and MR3 (Information Booths) as critical factors suggests that investments in signage, directions, and information services can enhance passenger experiences. Improving these aspects could lead to better navigation within the airport and higher overall satisfaction.
2. The lack of significant differences in factor scores across income levels, age groups, and gender suggests that airport services may need to adopt a universal approach to enhance passenger satisfaction. Service improvements should focus on general enhancements applicable to all demographic groups rather than targeting specific segments.
3. The significant differences in mean sentiment scores among different age groups indicate that age may influence passenger perceptions, suggesting tailored engagement strategies and services may be necessary to cater to varying preferences and expectations across age demographics.
4. The minimal impact of survey administration methods on sentiment scores suggests that airport resource allocation to one method over the other may be possible without sacrificing data quality.

Recommendations

1. Improve Signage and Directions: Enhance the clarity and visibility of signs and directions inside the airport and on roadways (related to MR1 and MR4). This will help passengers navigate more efficiently and reduce stress, improving overall satisfaction.
2. Enhance Amenities: Upgrade and diversify offerings in restaurants and retail shops. Consider expanding or enhancing artwork and exhibitions to improve the overall passenger experience, creating a more pleasant and engaging environment.
3. Optimize Parking and Transportation: Improve the efficiency and user-friendliness of airport parking facilities (related to MR2). Enhance shuttle services to the long-term parking lot and rental car center (related to MR5) to streamline transportation, making the airport more accessible and convenient for travelers.
4. Boost Information Services: Make information booths more accessible and user-friendly (related to MR3). Improve the clarity and availability of information on screens and monitors (related to MR1) to help passengers find the information they need more easily, greatly enhancing the customer service experience at the airport.
5. Tailored Communication and Services: Since age significantly influences sentiment, tailor communication and services to address the needs and preferences of different age groups. For example, provide more digital information for younger travelers and more in-person assistance for older travelers. Implement strategies that cater to the different needs and preferences of male and female travelers, ensuring that facilities and services are inclusive and address the unique concerns of each gender.

6. Enhance User Experience in Areas of Negative Sentiment: Conduct further qualitative analysis of the lower sentiment scores to identify specific pain points. Prioritize improvements in these areas to enhance user satisfaction.
7. Training Staff on Customer Engagement: Provide staff training focusing on understanding and responding to the varying sentiments of different age groups and genders. This training should emphasize personalized communication techniques to enhance the passenger experience.
8. Enhance Self-Administered Survey Options: Since self-administered surveys showed slightly higher sentiment scores, consider expanding self-service feedback options. Implement more kiosks or online platforms that allow passengers to provide feedback at their convenience, potentially increasing overall sentiment and response rates.
9. Conduct Follow-Up Studies: Conduct qualitative follow-up studies (e.g., focus groups) to delve deeper into the sentiments expressed by different age groups and genders. Understanding the context behind their feelings can lead to more effective interventions and improvements.

As the airport continues to evolve, fostering a culture of continuous improvement based on data-driven insights is essential. By implementing these recommendations, the airport can enhance the travel experience for all passengers, creating a more welcoming, efficient, and enjoyable environment. Engaging in ongoing dialogue with travelers and adapting to their changing needs will be key to sustaining high satisfaction levels and ensuring the airport remains a preferred choice for all who travel through it.