항공사 만족도에 영향을 미치는 요인 분석

-요약

코로나19가 장기화됨에 따라 해외를 가는 사람들이 급격히 줄어 항공사들은 위기에 직면하였고, 이에 항공사들 역시 고객들을 끌기 위한 대책을 마련하고 있다.

따라서 본 연구는 어떠한 변수들이 항공사 만족도에 영향을 미치는지 알아보고자 한다.

항공사 만족도와 연관이 있는 변수를 살펴보기 위해 독립성 검정을 실시하였고, 로지스틱 회귀분석을 통해 항공사 만족도에 영향을 주는 요인들을 파악했다. 또한 ROC곡선을 통해 모형의 적합성 역시 검토하였다. 위와 같은 분석을 통해 항공사 만족도에 영향을 미치는 요인들을 알아내어 개선점을 찾아내고자 한다.

-역할

201932119 응용통계학과 박나운 : ppt 제작 및 데이터 분석

201932144 응용통계학과 임정은 : 보고서 제작 및 데이터 분석

201932145 응용통계학과 임지연 : 발표 및 데이터 분석

목차

1. 서론

1.1 연구 배경 및 목적

갑작스럽게 나타난 코로나19로 인해 우리의 일상은 완전히 바뀌어 버렸고, 외출이 줄어든 만큼 항공사 역시 직접적인 피해에 직면해 왔었다. 그러나 최근 백신 접종이 완료되어 감에 따라 위드 코로나의 시대로 접해가고 있다.

이 시점에 항공사들은 그동안 항공사에 만족하지 못했던 고객들도 다시금 사로잡을 중요한 기회를 가지게 될 것이다. 그렇기에 평소 고객들이 항공사에 불만족하던 요인이 무엇인지 파악하는 과정이 필요하다.

따라서 본 연구는 고객들로 하여금 항공사 만족도에 영향을 미치는 요인이 무엇인지, 어떠한 점을 개선해야 할지 독립성 검정, 분할표분석, 상관행렬, 로지스틱 회귀분석 등의 방법으로 이를 분석하고자 한다.

2. 본론

2.1 데이터 설명

항공사 만족도에 영향을 미치는 요인을 분석하기 위해 kaggle에서 제공한 데이터를 활용하였다. 총 데이터는 25976개이며, 24개의 변수로 이루어져 있다. 이 중 분석에 불필요한 변수를 제외하고 20개의 변수를 남겨두었다.

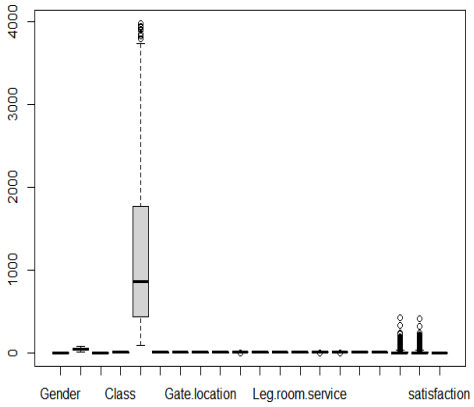
독립변수는 성별, 나이, 비행 목적, 비행기 클래스 등 19개로 항공사 만족도에 영향을 줄 가능성이 있는 설명 변수들이다. 종속변수는 항공사 만족도로 중립 또는 불만=0, 만족=1로 나타내었다.

	변수	**유형**	**비고**
설명 변수	Gender	명목형	승객의 성별(남성=1, 여성=2)
	Age	연속형	승객의 나이
	Type of Travel	명목형	비행 목적(출장=1, 개인여행=2)
	Class	순서형	비행기 클래스 (에코=1, 에코 플러스=2, 비즈니스=3)
	Flight Distance	연속형	비행 거리
	Inflight wifi service	순서형	기내 와이파이 서비스(0:적용불가)
	Departure/Arrival time convenient	순서형	출발/도착 시간의 편리성(0~5)
	Gate location	순서형	게이트 위치의 만족도(1~5)
	Food and drink	순서형	음식 및 음료의 만족도(0~5)
	Seat comfort	순서형	시트 편안함의 만족도(1~5)
	Inflight entertainment	순서형	기내 엔터테인먼트만족도(1~5)
	On-board service	순서형	기내 서비스 만족도(1~5)
	Leg room service	순서형	좌석간 거리 서비스의 만족도(0~5)
	Baggage handling	순서형	수하물 취급 만족도(1~5)
	Checkin service	순서형	체크인 서비스의 만족도(1~5)
	Inflight service	순서형	공항 내 항공사 서비스 만족도(1~5)
	Cleanliness	순서형	청결도의 만족도(1~5)
	Departure Delay in Minutes	연속형	출발 지연 시간(분)
	Arrival Delay in Minutes	연속형	도착 지연 시간(분)
종속 변수	satisfaction	명목형	항공사 만족도 (중립 또는 불만=0, 만족=1)

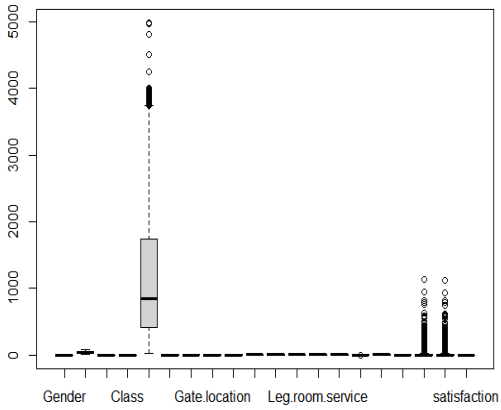
*만족도(1-5;매우불만족-매우만족)

```
> summary(air)
     Gender           Age        Type.of.Travel      Class       Flight.Distance Inflight.wifi.service Departure.Arrival.time.convenient
 Min.   :1.000   Min.   : 7.00   Min.   :1.000   Min.   :1.000   Min.   :  31    Min.   :0.000         Min.   :0.000
 1st Qu.:1.000   1st Qu.:27.00   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 414    1st Qu.:2.000         1st Qu.:2.000
 Median :2.000   Median :40.00   Median :1.000   Median :2.000   Median : 849    Median :3.000         Median :3.000
 Mean   :1.507   Mean   :39.62   Mean   :1.306   Mean   :2.036   Mean   :1194    Mean   :2.724         Mean   :3.046
 3rd Qu.:2.000   3rd Qu.:51.00   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:1744    3rd Qu.:4.000         3rd Qu.:4.000
 Max.   :2.000   Max.   :85.00   Max.   :2.000   Max.   :3.000   Max.   :4983    Max.   :5.000         Max.   :5.000
 Gate.location   Food.and.drink   Seat.comfort   Inflight.entertainment On.board.service Leg.room.service Baggage.handling Checkin.service
 Min.   :1.000   Min.   :0.000   Min.   :1.000   Min.   :0.000          Min.   :0.000    Min.   :0.00     Min.   :1.000    Min.   :1.000
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000          1st Qu.:2.000    1st Qu.:2.00     1st Qu.:3.000    1st Qu.:3.000
 Median :3.000   Median :3.000   Median :4.000   Median :4.000          Median :4.000    Median :4.00     Median :4.000    Median :3.000
 Mean   :2.976   Mean   :3.215   Mean   :3.449   Mean   :3.357          Mean   :3.386    Mean   :3.35     Mean   :3.633    Mean   :3.314
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000          3rd Qu.:4.000    3rd Qu.:4.00     3rd Qu.:5.000    3rd Qu.:4.000
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000          Max.   :5.000    Max.   :5.00     Max.   :5.000    Max.   :5.000
 Inflight.service  Cleanliness    Departure.Delay.in.Minutes Arrival.Delay.in.Minutes  satisfaction
 Min.   :0.000     Min.   :0.000   Min.   :   0.00           Min.   :   0.00          Min.   :0.0000
 1st Qu.:3.000     1st Qu.:2.000   1st Qu.:   0.00           1st Qu.:   0.00          1st Qu.:0.0000
 Median :4.000     Median :3.000   Median :   0.00           Median :   0.00          Median :0.0000
 Mean   :3.649     Mean   :3.286   Mean   :  14.23           Mean   :  14.74          Mean   :0.4389
 3rd Qu.:5.000     3rd Qu.:4.000   3rd Qu.:  12.00           3rd Qu.:  13.00          3rd Qu.:1.0000
 Max.   :5.000     Max.   :5.000   Max.   :1128.00           Max.   :1115.00          Max.   :1.0000
> summary(airplane2)
     Gender           Age        Type.of.Travel      Class       Flight.Distance Inflight.wifi.service Departure.Arrival.time.convenient
 Min.   :1.000   Min.   : 7.00   Min.   :1.000   Min.   :1.000   Min.   :  86    Min.   :0.000         Min.   :0.000
 1st Qu.:1.000   1st Qu.:27.00   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 431    1st Qu.:2.000         1st Qu.:2.000
 Median :1.000   Median :39.00   Median :1.000   Median :3.000   Median : 862    Median :3.000         Median :3.000
 Mean   :1.499   Mean   :39.59   Mean   :1.308   Mean   :2.076   Mean   :1232    Mean   :2.718         Mean   :3.074
 3rd Qu.:2.000   3rd Qu.:52.00   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:1769    3rd Qu.:4.000         3rd Qu.:4.000
 Max.   :2.000   Max.   :80.00   Max.   :2.000   Max.   :3.000   Max.   :3982    Max.   :5.000         Max.   :5.000
 Gate.location   Food.and.drink   Seat.comfort   Inflight.entertainment On.board.service Leg.room.service Baggage.handling Checkin.service
 Min.   :1.000   Min.   :0.000   Min.   :1.000   Min.   :1.000          Min.   :1.000    Min.   :0.000    Min.   :1.000    Min.   :1.000
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:2.000          1st Qu.:2.000    1st Qu.:2.000    1st Qu.:3.000    1st Qu.:3.000
 Median :3.000   Median :3.000   Median :4.000   Median :3.000          Median :4.000    Median :3.000    Median :4.000    Median :4.000
 Mean   :3.024   Mean   :3.113   Mean   :3.433   Mean   :3.237          Mean   :3.425    Mean   :3.286    Mean   :3.567    Mean   :3.386
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000          3rd Qu.:4.000    3rd Qu.:4.000    3rd Qu.:4.000    3rd Qu.:4.000
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000          Max.   :5.000    Max.   :5.000    Max.   :5.000    Max.   :5.000
 Inflight.service  Cleanliness    Departure.Delay.in.Minutes Arrival.Delay.in.Minutes  satisfaction
 Min.   :1.000     Min.   :1.000   Min.   :  0.00            Min.   :  0.00           Min.   :0.0000
 1st Qu.:3.000     1st Qu.:2.000   1st Qu.:  0.00            1st Qu.:  0.00           1st Qu.:0.0000
 Median :4.000     Median :3.000   Median :  0.00            Median :  0.00           Median :0.0000
 Mean   :3.581     Mean   :3.195   Mean   : 14.97            Mean   : 15.72           Mean   :0.4145
 3rd Qu.:5.000     3rd Qu.:4.000   3rd Qu.: 13.00            3rd Qu.: 14.00           3rd Qu.:1.0000
 Max.   :5.000     Max.   :5.000   Max.   :420.00            Max.   :407.00           Max.   :1.0000
```

airplane2 boxplot



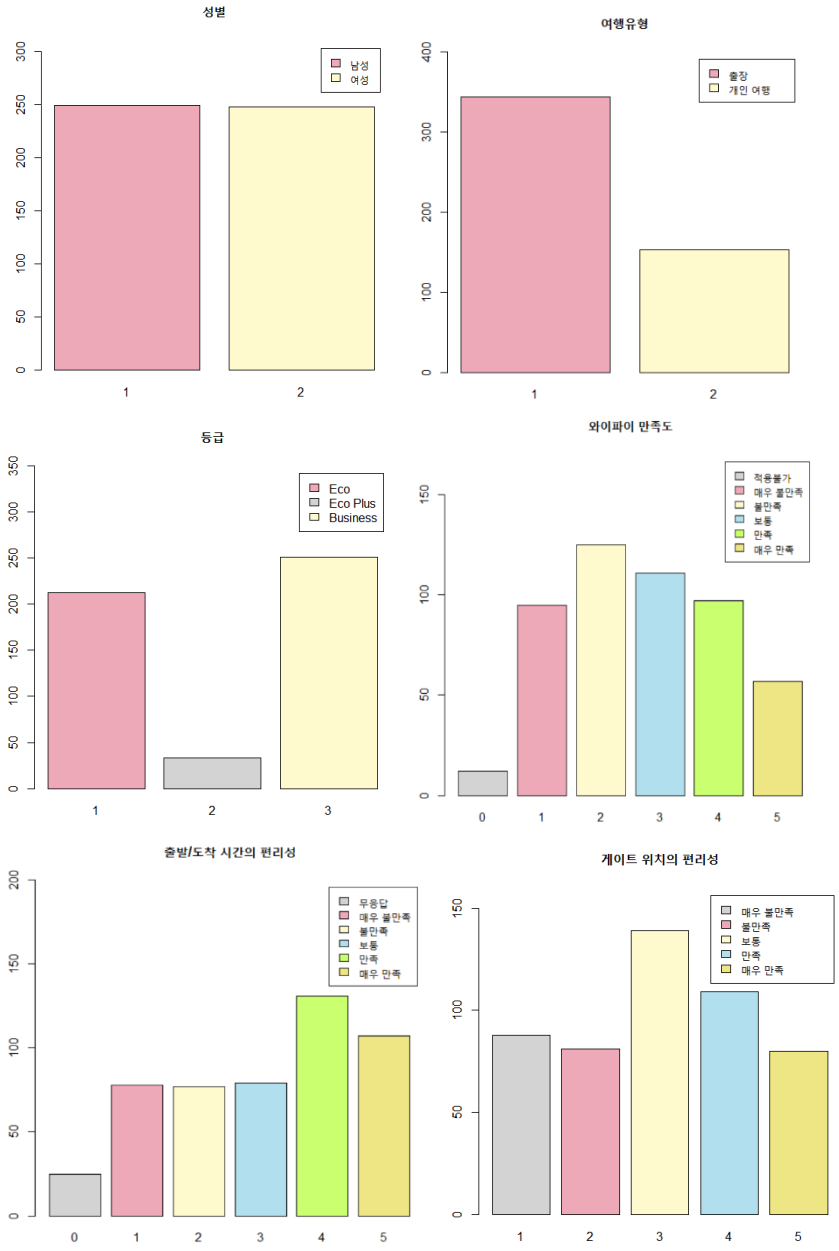
air boxplot



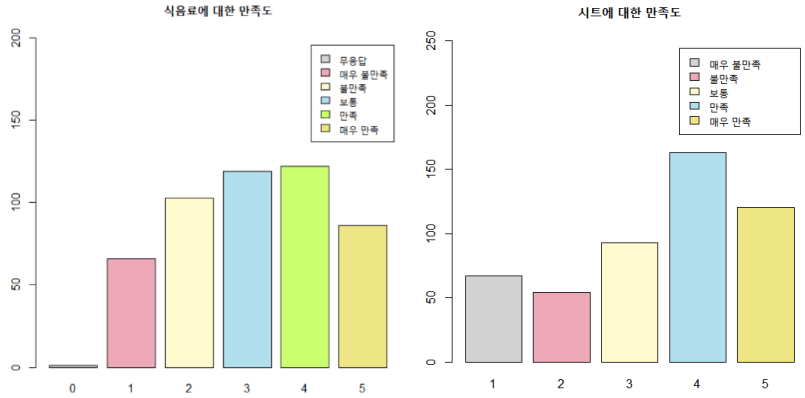
원자료인 air와 500개를 랜덤추출한 airplane2 자료의 통계량과 박스 플롯을 비교하였을 때 거의 차이가 나지 않음을 확인할 수 있다. 따라서 airplane2의 자료를 이용하여 분석을 실시하였다.

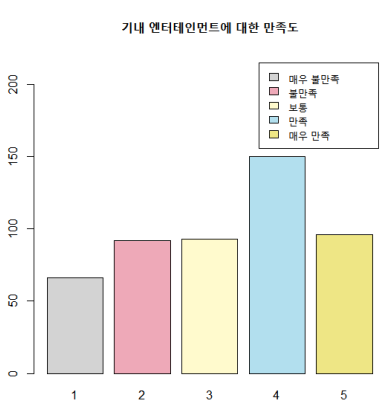
2.2 EDA

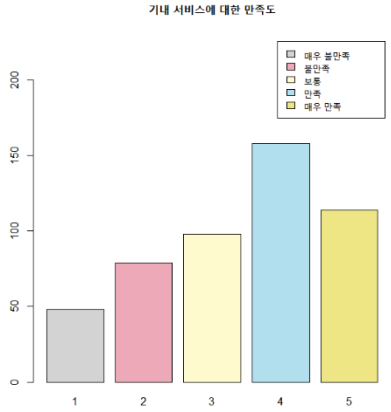
데이터 분석에 앞서 개별 변수들의 특성을 파악하기 위해 데이터를 시각화하여 나타내었다.

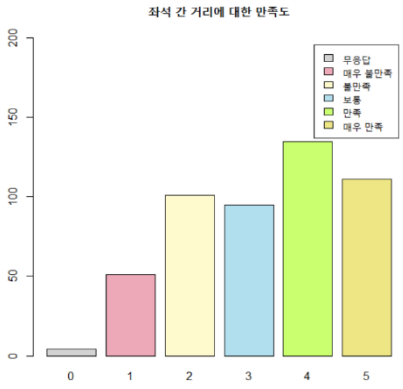


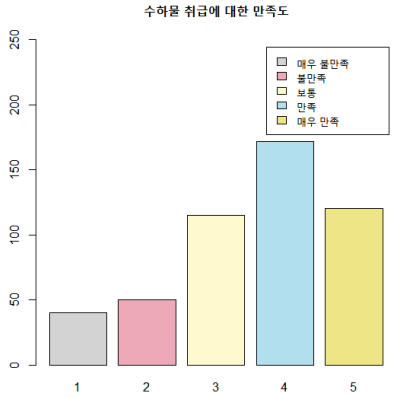
응답자의 성별 비율에는 거의 차이가 없음을 알 수 있으며, 여행 유형에 관해서는 출장인 경우가 개인 여행에 비해 약 2배정도 많다는 것을 확인할 수 있다. 비행기 등급에 대해서는 비즈니스석인 경우가 가장 많았고, 에코 플러스인 경우는 가장 적었다. 와이파이 만족도에 대해서는 보통이거나 만족이라는 답변이 많았고, 매우 불만족이라는 응답은 매우 적었다. 출발, 도착 시간의 편리성에 대해서는 만족한다는 응답이 많았고, 매우 불만족, 불만족, 보통인 비율이 비슷하였다. 게이트의 위치에 대해서는 보통이라는 응답이 가장 많았다.



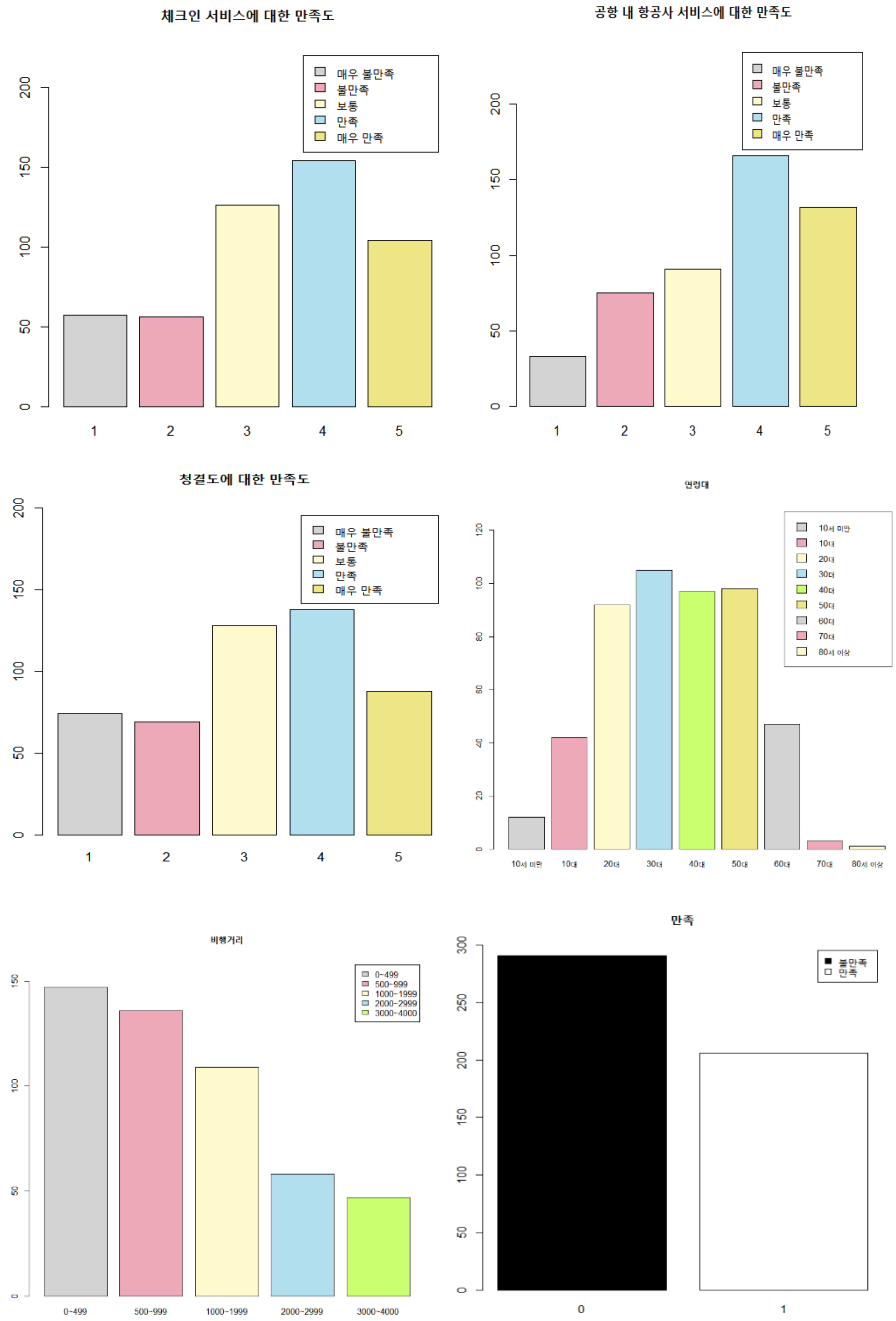








식음료에 대한 만족도에 대해서는 만족과 보통이라는 응답이 가장 많았고, 시트에 대해서도 만족이라는 응답이 가장 많았다. 기내 엔터테인먼트에 대해서는 만족이라는 응답이 가장 많았고, 나머지 응답은 비슷하게 나타났다. 기내 서비스에 대해서는 대체로 만족한다는 비율이 높았고, 좌석 간 거리에 대해서는 만족이라는 응답이 가장 많았지만, 불만족, 보통, 매우 만족인 비율이 비슷하게 나타났다. 수하물 취급에 대해서는 만족한다는 응답이 가장 많았고, 보통과 매우 만족의 응답 비율이 비슷하게 높게 나타났다.



체크인 서비스, 공항 내 항공사 서비스, 청결도에 대해서 모두 만족한다는 응답이 가장 많았고, 항공사 연령대는 30 대가 가장 많았지만, 20 대, 30 대, 50 대가 모두 비슷하게 높게 나타났다. 비행거리는 0~499 인 단거리 비행이 가장 많았고, 거리가 멀수록 작아지는 경향이 있다. 종속변수인 항공사 만족도는 만족보다 불만족이라는 답변이 더 많았다. 어떤 요인으로 인해 항공사 만족도에 불만족이라는 답변이 많았는지 다음 분석을 통해 자세히 살펴보고자 한다.

2.3 독립성검정

로지스틱 회귀분석을 실시하기 전에 범주형 자료에서 각 변수들과 항공사 만족도 간의 연관성을 판단하기 위해 독립성 검정을 실시하였다.

독립성 검정 결과

(1)명목형 중에서

```
> chisq.test(airplane2$satisfaction, airplane2$Gender)

        Pearson's Chi-squared test with Yates' continuity correction

data:  airplane2$satisfaction and airplane2$Gender
X-squared = 0.0028423, df = 1, p-value = 0.9575

> chisq.test(airplane2$satisfaction, airplane2$Type.of.Travel)

        Pearson's Chi-squared test with Yates' continuity correction

data:  airplane2$satisfaction and airplane2$Type.of.Travel
X-squared = 96.95, df = 1, p-value < 2.2e-16
```

Chi-squared test에서 Gender(성별)은 p-value가 0.9575로 귀무가설(H0: 서로 독립이다)을 기각할 수 없다. 따라서 항공사 만족도와 성별은 관련이 있다고 볼 수 없다.
Chi-squared test에서 type of travel(여행 유형)이 p-value가 2.2e-16으로 매우 작기 때문에 귀무가설(H0: 서로 독립이다)을 기각한다. 따라서 항공사 만족도와 type of travel(여행 유형)은 관련 있을 것이다.

(2)순서형 중에서

Class

Class	항공사 만족도	
	불만족(0)	만족(1)
1(에코)	174	39
2(에코 플러스)	27	6
3(비즈니스)	90	161

```
> CMHtest(class,recores=c(1,2,3,4,5))
Cochran-Mantel-Haenszel Statistics for Class by satisfaction

                AltHypothesis  Chisq Df       Prob
cor         Nonzero correlation 100.68  1 1.0813e-23
rmeans   Row mean scores differ 107.40  2 4.7570e-24
cmeans   Col mean scores differ 100.68  1 1.0813e-23
general     General association 107.40  2 4.7570e-24
```

CMH test에서 Class(비행기 클래스)는 p-value가 1.0813e-23으로 매우 작으므로 귀무가설(H0: 서로 독립이다)을 기각한다. 따라서 비행기 클래스와 항공사 만족도는 관련이 있을 것이다.

Inflight wifi service

Inflight wifi service	항공사 만족도	
	불만족(0)	만족(1)
0(적용불가)	0	12
1(매우 불만족)	66	29
2(불만족)	99	26
3(보통)	82	29
4(만족)	44	53
5(매우 만족)	0	57

```
> CMHtest(inflight.wifi.service,recores=c(1,2,3,4,5))
Cochran-Mantel-Haenszel Statistics for Inflight.wifi.service by satisfaction

                AltHypothesis   Chisq Df       Prob
cor        Nonzero correlation  50.482  1 1.2023e-12
rmeans  Row mean scores differ 141.509  5 8.5490e-29
cmeans  Col mean scores differ  50.482  1 1.2023e-12
general    General association 141.509  5 8.5490e-29
```

CMH test 에서 Inflight wife service(기내 와이파이 서비스)는 p-value 가 1.2023e-12 로 매우 작기 때문에 귀무가설(H0:서로 독립이다)을 기각한다. 따라서 기내 와이파이 서비스와 항공사 만족도는 관련이 있을 것이다.

Departure/Arrival time convenient

Departure/Arrival time convenient	항공사 만족도	
	불만족(0)	만족(1)
0()	16	9
1(매우 불만족)	41	37
2(불만족)	50	27
3(보통)	41	38
4(만족)	78	53
5(매우 만족)	65	42

```
> CMHtest(Departure,recores=c(1,2,3,4,5))
Cochran-Mantel-Haenszel Statistics for Departure.Arrival.time.convenient by satisfaction

                AltHypothesis   Chisq Df    Prob
cor        Nonzero correlation 0.14107  1 0.70722
rmeans  Row mean scores differ 4.44828  5 0.48683
cmeans  Col mean scores differ 0.14107  1 0.70722
general    General association 4.44828  5 0.48683
```

CMH test 에서 Departure/Arrival time convenient(출발/도착 시간의 편리성)은 p-value 가 0.70722 로 유의수준 0.70722 로 귀무가설(H0: 서로 독립이다)을 기각할 수 없다. 따라서 출발/도착 시간의 편리성과 항공사 만족도는 관련이 있다고 볼 수 없다.

Gate location

Gate location	항공사 만족도	
	불만족(0)	만족(1)
1(매우 불만족)	49	39
2(불만족)	44	37
3(보통)	94	45
4(만족)	68	41
5(매우 만족)	36	44

```
> CMHtest(Gate,recores=c(1,2,3,4,5))
Cochran-Mantel-Haenszel Statistics for Gate.location by satisfaction

                AltHypothesis   Chisq Df     Prob
cor        Nonzero correlation  0.3888  1 0.532933
rmeans  Row mean scores differ 12.3013  4 0.015246
cmeans  Col mean scores differ  0.3888  1 0.532933
general    General association 12.3013  4 0.015246
```

CMH test 에서 Gate location(게이트 위치)은 p-value 가 0.532933 으로 0.001 보다 크기 때문에 귀무가설(H0: 서로 독립이다)을 기각할 수 없다. 따라서 게이트 위치와 항공사 만족도는 관련이 있다고 볼 수 없다.

Food and drink

Food and drink	항공사 만족도	
	불만족(0)	만족(1)
0()	1	0
1(매우 불만족)	53	13
2(불만족)	57	46
3(보통)	68	51
4(만족)	70	52
5(매우 만족)	42	44

```
> CMHtest(Food,recores=c(1,2,3,4,5))
Cochran-Mantel-Haenszel Statistics for Food.and.drink by satisfaction

                AltHypothesis   Chisq Df      Prob
cor        Nonzero correlation  9.8712  1 0.0016789
rmeans  Row mean scores differ 17.4881  5 0.0036614
cmeans  Col mean scores differ  9.8712  1 0.0016789
general    General association 17.4881  5 0.0036614
```

CMH test 에서 Food and drink(식음료)의 p-value 가 0.0016789 로 0.001 보다 크기 때문에 귀무가설(H0: 서로 독립이다)을 기각할 수 없다. 따라서 식음료와 항공사 만족도는 관련이 있다고 볼 수 없다.

Seat comfort

Seat comfort	항공사 만족도	
	불만족(0)	만족(1)
1(매우 불만족)	55	12
2(불만족)	39	15
3(보통)	66	27
4(만족)	86	77
5(매우 만족)	45	75

```
> CMHtest(Seat,recores=c(1,2,3,4,5))
Cochran-Mantel-Haenszel Statistics for Seat.comfort by satisfaction

                AltHypothesis  Chisq Df       Prob
cor         Nonzero correlation 46.114  1 1.1159e-11
rmeans   Row mean scores differ 49.427  4 4.7553e-10
cmeans   Col mean scores differ 46.114  1 1.1159e-11
general     General association 49.427  4 4.7553e-10
```

CMH test 에서 Seat comfort(좌석의 편안함)의 p-value 가 1.1159e-11 로 매우 작기 때문에 귀무가설(H0: 서로 독립이다)을 기각한다. 따라서 좌석의 편안함과 항공사 만족도는 관련이 있다고 볼 수 있다.

Inflight entertainment

Inflight entertainment	항공사 만족도	
	불만족(0)	만족(1)
1(매우 불만족)	58	8
2(불만족)	73	19
3(보통)	65	28
4(만족)	58	92
5(매우 만족)	37	59

```
> CMHtest(Inflight,recores=c(1,2,3,4,5))
Cochran-Mantel-Haenszel Statistics for Inflight.entertainment by satisfaction

                AltHypothesis  Chisq Df       Prob
cor         Nonzero correlation 75.848  1 3.0637e-18
rmeans   Row mean scores differ 84.821  4 1.6556e-17
cmeans   Col mean scores differ 75.848  1 3.0637e-18
general     General association 84.821  4 1.6556e-17
```

CMH test 에서 Inflight entertainment(기내 엔터테인먼트)의 p-value 가 3.0637e-18 로 매우 작기 때문에 귀무가설(H0: 서로 독립이다)을 기각한다. 따라서 기내 엔터테인먼트와 항공사 만족도는 관련이 있다고 볼 수 있다.

On-board service

On-board service	항공사 만족도	
	불만족(0)	만족(1)
1(매우 불만족)	40	8
2(불만족)	61	18
3(보통)	69	29
4(만족)	81	77
5(매우 만족)	40	74

```
> CMHtest(On.board,recores=c(1,2,3,4,5))
Cochran-Mantel-Haenszel Statistics for On.board.service by satisfaction

                AltHypothesis  Chisq Df       Prob
cor         Nonzero correlation 55.427  1 9.6974e-14
rmeans   Row mean scores differ 58.362  4 6.4054e-12
cmeans   Col mean scores differ 55.427  1 9.6974e-14
general     General association 58.362  4 6.4054e-12
```

CMH test 에서 On-board service(기내 서비스)의 p-value 가 9.6974e-14 로 매우 작기 때문에 귀무가설(H0: 서로 독립이다)을 기각한다. 따라서 기내 서비스와 항공사 만족도는 관련이 있다고 볼 수 있다.

Leg room service

leg room service	항공사 만족도	
	불만족(0)	만족(1)
0(	1	3
1(매우 불만족)	43	8
2(불만족)	77	24
3(보통)	72	23
4(만족)	59	76
5(매우 만족)	39	72

```
> CMHtest(Leg,recores=c(1,2,3,4,5))
Cochran-Mantel-Haenszel Statistics for Leg.room.service by satisfaction

                AltHypothesis  Chisq Df       Prob
cor         Nonzero correlation 58.806  1 1.7402e-14
rmeans   Row mean scores differ 77.638  5 2.6159e-15
cmeans   Col mean scores differ 58.806  1 1.7402e-14
general     General association 77.638  5 2.6159e-15
```

CMHtest 에서 Leg room service(좌석 간 거리)의 p-value 가 1.7402e-14 로 매우 작기 때문에 귀무가설(H0: 서로 독립이다)을 기각한다. 따라서 좌석 간 거리와 항공사 만족도는 관련이 있다고 볼 수 있다.

Baggage handling

Baggage handling	항공사 만족도	
	불만족(0)	만족(1)
1(매우 불만족)	33	7
2(불만족)	35	15
3(보통)	86	29
4(만족)	85	87
5(매우 만족)	52	68

```
> CMHtest(Baggage,recores=c(1,2,3,4,5))
Cochran-Mantel-Haenszel Statistics for Baggage.handling by satisfaction

                AltHypothesis  Chisq Df       Prob
cor         Nonzero correlation 34.835  1 3.5880e-09
rmeans   Row mean scores differ 41.916  4 1.7368e-08
cmeans   Col mean scores differ 34.835  1 3.5880e-09
general     General association 41.916  4 1.7368e-08
```

CMH test 에서 Baggage handling(수하물 취급)의 p-value 가 3.5880e-09 로 매우 작기 때문에 귀무가설(H0: 서로 독립이다)을 기각한다. 따라서 수하물 취급과 항공사 만족도는 관련이 있다고 볼 수 있다.

Checkin service

Checkin service	항공사 만족도	
	불만족(0)	만족(1)
1(매우 불만족)	40	17
2(불만족)	39	17
3(보통)	75	51
4(만족)	83	71
5(매우 만족)	54	50

```
> CMHtest(Checkin,recores=c(1,2,3,4,5))
Cochran-Mantel-Haenszel Statistics for Checkin.service by satisfaction

                AltHypothesis  Chisq Df      Prob
cor         Nonzero correlation 8.6147  1 0.0033346
rmeans   Row mean scores differ 9.3005  4 0.0540117
cmeans   Col mean scores differ 8.6147  1 0.0033346
general     General association 9.3005  4 0.0540117
```

CMH test 에서 Checkin service(체크인 서비스)의 p-value 가 0.0033346 으로 0.001 보다 크기 때문에 귀무가설(H0: 서로 독립이다)을 기각할 수 없다. 따라서 체크인 서비스와 항공사 만족도는 관련이 있다고 볼 수 없다.

Inflight service

Inflight service	항공사 만족도	
	불만족(0)	만족(1)
1(매우 불만족)	28	5
2(불만족)	56	19
3(보통)	66	25
4(만족)	83	83
5(매우 만족)	58	74

```
> CMHtest(service,recores=c(1,2,3,4,5))
Cochran-Mantel-Haenszel Statistics for Inflight.service by satisfaction

                AltHypothesis  Chisq Df       Prob
cor         Nonzero correlation 37.922  1 7.3642e-10
rmeans   Row mean scores differ 41.285  4 2.3462e-08
cmeans   Col mean scores differ 37.922  1 7.3642e-10
general     General association 41.285  4 2.3462e-08
```

CMH test 에서 Inflight service(공항 내 항공사 서비스)의 p-value 가 7.3642e-10 으로 매우 작기 때문에 귀무가설(H0: 서로 독립이다)을 기각하므로 공항 내 항공사 서비스와 항공사 만족도는 관련이 있다고 볼 수 있다.

Cleanliness

Cleanliness	항공사 만족도	
	불만족(0)	만족(1)
1(매우 불만족)	60	14
2(불만족)	55	14
3(보통)	65	63
4(만족)	73	65
5(매우 만족)	38	50

```
> CMHtest(clean,recores=c(1,2,3,4,5))
Cochran-Mantel-Haenszel Statistics for Cleanliness by satisfaction

                AltHypothesis  Chisq Df       Prob
cor         Nonzero correlation 33.716  1 6.3759e-09
rmeans   Row mean scores differ 41.689  4 1.9352e-08
cmeans   Col mean scores differ 33.716  1 6.3759e-09
general     General association 41.689  4 1.9352e-08
```

CMH test 에서 Cleanliness(청결도)의 p-value 가 6.3759e-09 로 매우 작기 때문에 귀무가설(H0: 서로 독립이다)을 기각한다. 따라서 청결도와 항공사 만족도는 관련이 있다고 볼 수 있다.

명목형 변수 중 Gender(성별), 순서형 변수 중 Departure/Arrival time convenient(출발/도착 시간의 편리성), Gate location(게이트 위치), Food and drink(식음료) , Chickin service(체크인 서비스)는 p-value 가 0.001 보다 크므로 귀무가설(H0: 서로 독립이다)을 기각할 수 없다. 따라서

satisfaction(항공사 만족도)와 출발/도착 편리성, 게이트 위치, 식음료, 체크인 서비스는 독립이고, 나머지 변수들은 항공사 서비스와 관련이 있다고 볼 수 있다.

2.4 상관행렬

모형 적합시 다중공선성의 문제가 발생할 것을 대비하여 연속형 변수들 간의 상관관계를 확인하였다. Departure.Delay.in.Minutes(출발 지연 시간)과 Arrival.Delay.in.Minutes(도착 지연 시간)의 상관계수가 0.96 으로 가장 크므로 강한 상관관계를 갖는다. 따라서 다중공선성의 문제가 발생할 수 있기 때문에 Arrival.Delay.in.Minutes(도착 지연 시간)으로 대신하여 회귀분석을 실시한다.

-Stepwise 방법을 통해 변수 축소

```
> fit<-glm(satisfaction ~ Age + factor(Type.of.Travel) + Class +  Flight.Distance + Inflight.wifi.service +
+           Seat.comfort + Inflight.entertainment + On.board.service + Leg.room.service +
+            Baggage.handling + Inflight.service + Cleanliness + Departure.Delay.in.Minutes, family = binomial, data = airplane2)
> summary(fit)

Call:
glm(formula = satisfaction ~ Age + factor(Type.of.Travel) + Class +
    Flight.Distance + Inflight.wifi.service + Seat.comfort +
    Inflight.entertainment + On.board.service + Leg.room.service +
    Baggage.handling + Inflight.service + Cleanliness + Departure.Delay.in.Minutes,
    family = binomial, data = airplane2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5281  -0.5718  -0.1887   0.5597   2.6997

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -9.7903247  1.0853780  -9.020  < 2e-16 ***
Age                         0.0361192  0.0100412   3.597 0.000322 ***
factor(Type.of.Travel)2    -2.3043187  0.4290411  -5.371 7.84e-08 ***
Class                       0.6965770  0.1695162   4.109 3.97e-05 ***
Flight.Distance             0.0003196  0.0001398   2.287 0.022210 *
Inflight.wifi.service       0.5780433  0.1094917   5.279 1.30e-07 ***
Seat.comfort                0.1963079  0.1463878   1.341 0.179916
Inflight.entertainment     -0.1248969  0.1807907  -0.691 0.489668
On.board.service            0.4676126  0.1426140   3.279 0.001042 **
Leg.room.service            0.0687421  0.1171259   0.587 0.557266
Baggage.handling            0.2514954  0.1576537   1.595 0.110659
Inflight.service            0.1070816  0.1716187   0.624 0.532660
Cleanliness                 0.4215706  0.1556410   2.709 0.006757 **
Departure.Delay.in.Minutes -0.0028224  0.0049331  -0.572 0.567231
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 661.18  on 496  degrees of freedom
Residual deviance: 370.70  on 483  degrees of freedom
AIC: 398.7

Number of Fisher Scoring iterations: 6
```

회귀식 도출하기 위해 원래의 자료를 이용하여 회귀식을 도출하였다.

Stepwise 방법 통해 유의수준 0.05 보다 큰 변수들을 제외하여 fit 에서 fit1 으로 변수를 추려내었다.

```
> fit1<-glm(formula = satisfaction ~ Age + factor(Type.of.Travel) + Class +
+             Flight.Distance + Inflight.wifi.service + On.board.service +
+             Baggage.handling + Cleanliness, family = binomial, data = airplane2)
> summary(fit1)

Call:
glm(formula = satisfaction ~ Age + factor(Type.of.Travel) + Class +
    Flight.Distance + Inflight.wifi.service + On.board.service +
    Baggage.handling + Cleanliness, family = binomial, data = airplane2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4677  -0.5634  -0.1982   0.5774   2.7257

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -9.4699977  1.0074600  -9.400  < 2e-16 ***
Age                       0.0362441  0.0096919   3.740 0.000184 ***
factor(Type.of.Travel)2 -2.2094334  0.4156813  -5.315 1.07e-07 ***
Class                     0.7616152  0.1651311   4.612 3.98e-06 ***
Flight.Distance           0.0003161  0.0001365   2.315 0.020587 *
Inflight.wifi.service     0.5917370  0.1074496   5.507 3.65e-08 ***
On.board.service          0.4857308  0.1255825   3.868 0.000110 ***
Baggage.handling          0.2713777  0.1372933   1.977 0.048084 *
Cleanliness               0.4782986  0.1040555   4.597 4.30e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 661.18  on 496  degrees of freedom
Residual deviance: 373.55  on 488  degrees of freedom
AIC: 391.55

Number of Fisher Scoring iterations: 6
```

```
> anova(fit1,fit)
Analysis of Deviance Table

Model 1: satisfaction ~ Age + factor(Type.of.Travel) + Class + Flight.Distance +
    Inflight.wifi.service + On.board.service + Baggage.handling +
    Cleanliness
Model 2: satisfaction ~ Age + factor(Type.of.Travel) + Class + Flight.Distance +
    Inflight.wifi.service + Seat.comfort + Inflight.entertainment +
    On.board.service + Leg.room.service + Baggage.handling +
    Inflight.service + Cleanliness + Departure.Delay.in.Minutes
  Resid. Df Resid. Dev Df Deviance
1       488     373.55
2       483     370.70  5    2.845
> AIC(fit,fit1)
     df      AIC
fit  14 398.7004
fit1  9 391.5454
```

Anova 와 AIC 로 두 모형 비교하여 방법의 정당함을 증명하였다.

AIC 가 제일 작은 것을 최선의 모형으로 선택하므로 fit1 의 AIC 가 더 작기 때문에 fit1 선택한다.

2.5 로지스틱 회귀분석

```
> fit2<-glm(formula = satisfaction ~ Age + factor(Type.of.Travel) + Class +
+                 Flight.Distance + Inflight.wifi.service + On.board.service +
+                 Baggage.handling + Cleanliness, family = binomial("logit"), data = airplane2)
> summary(fit2)

Call:
glm(formula = satisfaction ~ Age + factor(Type.of.Travel) + Class +
    Flight.Distance + Inflight.wifi.service + On.board.service +
    Baggage.handling + Cleanliness, family = binomial("logit"),
    data = airplane2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4677  -0.5634  -0.1982   0.5774   2.7257

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -9.4699977  1.0074600  -9.400  < 2e-16 ***
Age                      0.0362441  0.0096919   3.740 0.000184 ***
factor(Type.of.Travel)2 -2.2094334  0.4156813  -5.315 1.07e-07 ***
Class                    0.7616152  0.1651311   4.612 3.98e-06 ***
Flight.Distance          0.0003161  0.0001365   2.315 0.020587 *
Inflight.wifi.service    0.5917370  0.1074496   5.507 3.65e-08 ***
On.board.service         0.4857308  0.1255825   3.868 0.000110 ***
Baggage.handling         0.2713777  0.1372933   1.977 0.048084 *
Cleanliness              0.4782986  0.1040555   4.597 4.30e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 661.18  on 496  degrees of freedom
Residual deviance: 373.55  on 488  degrees of freedom
AIC: 391.55

Number of Fisher Scoring iterations: 6
```

모든 변수가 유의수준 0.05에서 유의하여 최종 모형으로 채택하였다.

최종 회귀모형

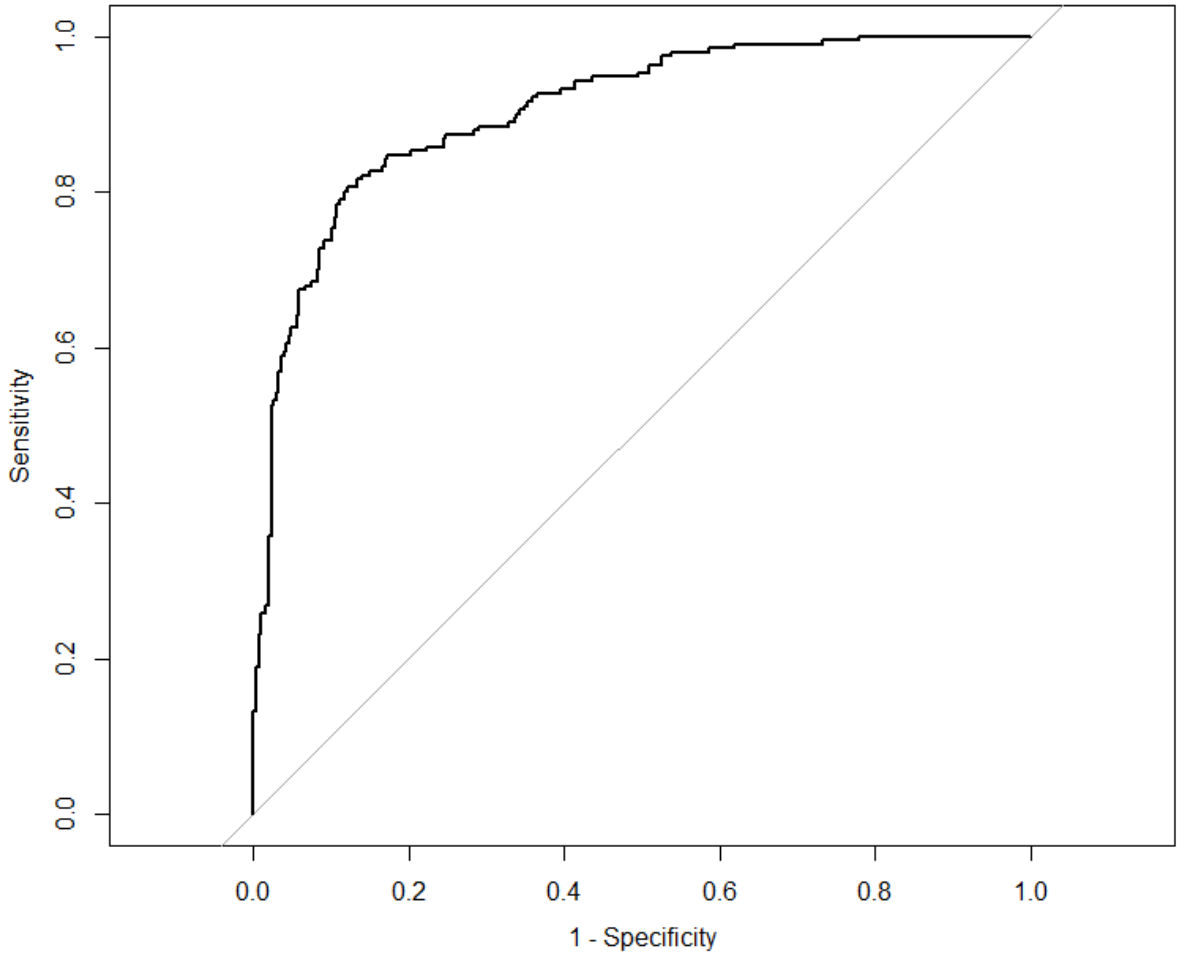
$$logit(\pi(x))$$
$$= -9.4699977 + 0.0362441(Age) - 2.2094334(factor(Type.of.Travel)2 + 0.7616252(Class)$$
$$+ 0.0003161(Flight.Distance) + 0.5917370(Inflight.wifi.service) + 0.4857308(On.board.service)$$
$$+ 0.2713777(Baggage.handling) + 0.4782986(Cleanliness)$$

-모형의 적합성 평가

```
> #deviance값을 통해 모형의 적합성 평가
> 1-pchisq(661.18-373.55,496-488)
[1] 0
```

deviance값을 이용하여 카이제곱 검정을 해본 결과, p-value가 0으로 유의하여 모형이 데이터에 잘 적합 한다는 것을 알 수 있다.

-ROC곡선



```
> auc(rocplot)
Area under the curve: 0.905
```

ROC 곡선을 이용하여 최종 모형에서의 예측 검정력을 확인해 보았다. ROC 커브의 아래면적이 넓을수록 성능이 좋다고 볼 수 있으므로 위 그림을 통해 모형이 잘 적합 되었다고 평가할 수 있다. 또한, AUC값은 90%로 이 모형의 예측 검정력은 높다고 할 수 있다.

-적합된 데이터의 해석

```
> exp(0.0362441)
[1] 1.036909
> exp(-2.2094334)
[1] 0.1097628
> exp(0.7616152)
[1] 2.141733
> exp(0.0003161)
[1] 1.000316
> exp(0.5917370)
[1] 1.807125
> exp(0.4857308)
[1] 1.625362
> exp(0.2713777)
[1] 1.31177
> exp(0.4782986)
[1] 1.613327
```

나이가 많을수록 항공사 만족도의 오즈가 1.04배 증가한다.

여행유형이 개인여행인 경우 항공사 만족도의 오즈는 여행유형이 출장인 경우의 항공사 만족도 오즈의 0.11배이다..

비행기 클래스가 높아질수록 항공사 만족도의 오즈가 2.14배 증가한다.

비행거리가 증가할수록 항공사 만족도의 오즈가 약 1배 증가한다.

기내 와이파이 서비스가 높아질수록 항공사 만족도의 오즈가 1.81배 증가한다.

기내 서비스가 높아질수록 항공사 만족도의 오즈가 1.63배 증가한다.

수화물 취급 만족도가 높아질수록 항공사 만족도의 오즈가 1.31배 증가한다.

청결도가 높아질수록 항공사 만족도의 오즈가 1.61배 증가한다.

3. 결론

최종적으로 설정된 로지스틱 회귀모형의 계수들을 살펴보면, 나이, 여행유형, 비행기 클래스, 비행거리, 기내 와이파이 서비스, 기내 서비스, 수하물 취급 만족도, 청결도가 항공사 만족도에 영향을 주는 대표적인 요인들임을 확인할 수 있다. 그 중 비행기 클래스가 에코인지, 에코플러스인지, 비즈니스인지에 따라 항공사 만족도에 가장 크게 영향을 주는 것으로 나타났다. 여행 유형과 같이 항공사에서 해결할 수 없는 것을 제외하고는 기내 와이파이 서비스, 기내 서비스, 수하물 취급 만족도, 청결도의 개선이 필요하며, 비행기 클래스에 관계없이 모든 고객들이 만족할 수 있도록 조금 더 세심한 주의가 필요하다.

독립성 검정결과, 성별, 출발/도착 시간의 편리성, 게이트 위치, 식음료, 체크인 서비스는 항공사 만족도와 관련이 없는 것으로 나타났고, 여행유형, 비행기 클래스, 기내 와이파이 서비스, 좌석의 편안함, 기내 엔터테인먼트 서비스, 기내 서비스, 좌석 간 거리, 수하물 취급, 공항 내 항공사 서비스, 청결도는 항공사 만족도와 관련이 있는 것으로 나타났다.

코로나19로 인해 해외를 가는 사람들이 줄어듦에 따라 항공사들 역시 막대한 피해를 입어왔다. 그러나 최근 위드 코로나로 전환되고 있는 이 시점이 항공사들에게 매우 중요하게 다가올 것이다. 고객들로 하여금 항공사 만족도에 영향을 미치는 요인들을 점검하고 이를 개선하여 코로나 걱정 없이 자유롭게 해외를 다니게 될 고객들의 마음을 더욱 사로잡을 수 있는 기회가 되기를 바라며, 우리 모두가 빠른 시일 내에 일상을 돌아갈 수 있기를 희망한다.

<출처>

https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction