# 2주차 EDA

김부현

# 1. 분석 데이터셋

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

1. 인구사회학적 요인

*Gender:* Gender of the passengers (Female, Male)
*Age:* The actual age of the passengers

2. 승객 유형 분류
*Customer Type:* The customer type (Loyal customer, disloyal customer)
*Type of Travel:* Purpose of the flight of the passengers (Personal Travel, Business Travel)
*Class:* Travel class in the plane of the passengers (Business, Eco, Eco Plus)
*Flight distance:* The flight distance of this journey

# 1. 분석 데이터셋

3. 개별 서비스 만족도

*Inflight wifi service:* Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
*Departure/Arrival time convenient:* Satisfaction level of Departure/Arrival time convenient
*Ease of Online booking:* Satisfaction level of online booking
*Gate location:* Satisfaction level of Gate location
*Food and drink:* Satisfaction level of Food and drink
*Online boarding:* Satisfaction level of online boarding
*Seat comfort:* Satisfaction level of Seat comfort
*Inflight entertainment:* Satisfaction level of inflight entertainment
*On-board service:* Satisfaction level of On-board service
*Leg room service:* Satisfaction level of Leg room service
*Baggage handling:* Satisfaction level of baggage handling
*Check-in service:* Satisfaction level of Check-in service
*Inflight service:* Satisfaction level of inflight service
*Cleanliness:* Satisfaction level of Cleanliness

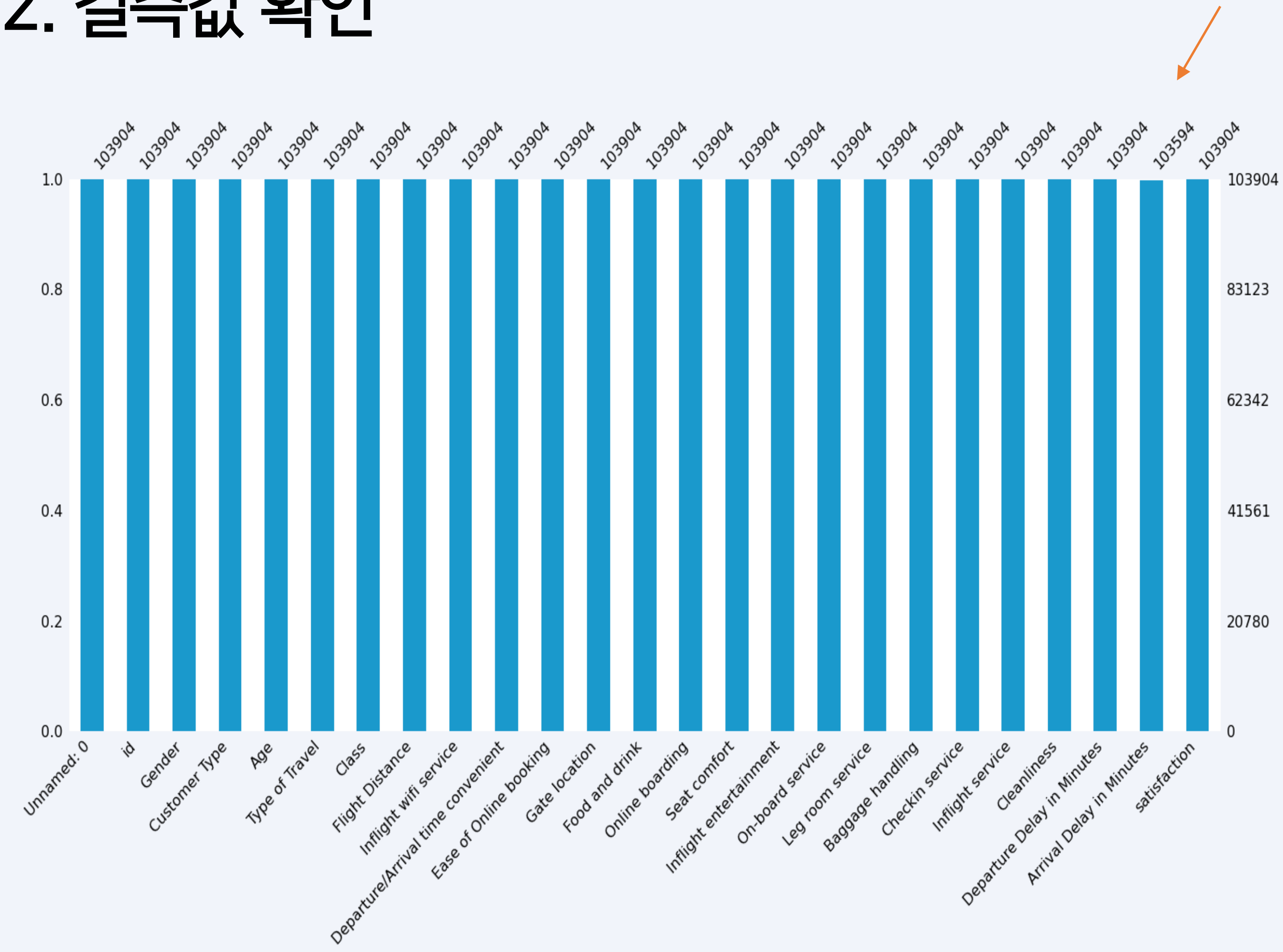
4. 지연시간

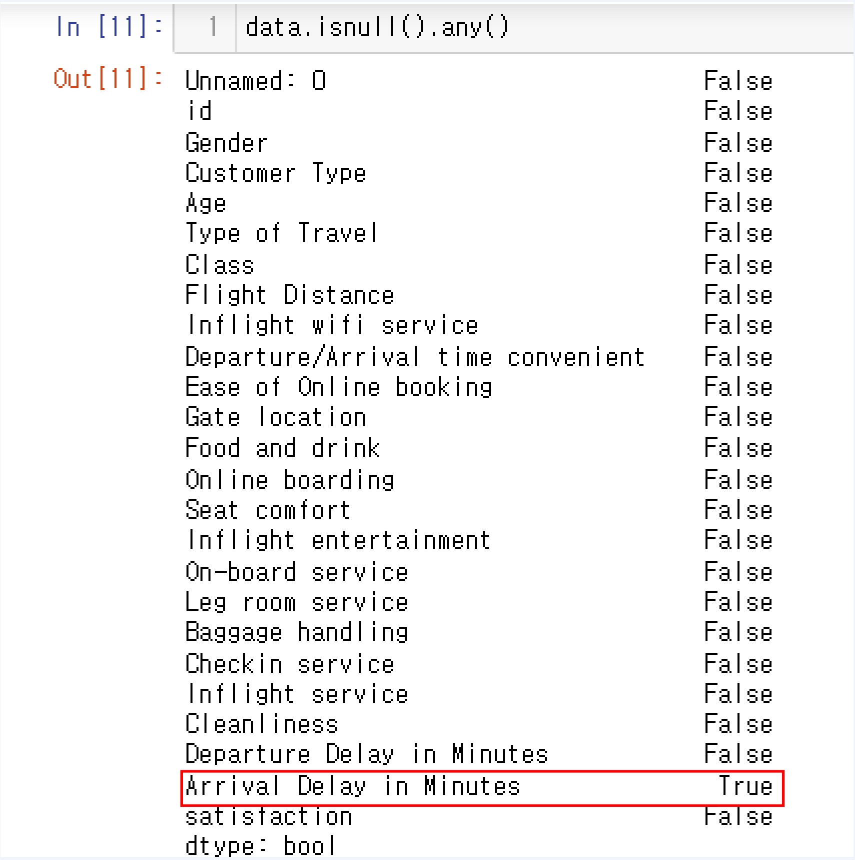
*Departure Delay in Minutes:* Minutes delayed when departure
*Arrival Delay in Minutes:* Minutes delayed when Arrival

5. 항공 만족도 (결과변수 - 이진형)

*Satisfaction:* Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

# 2. 결측값 확인



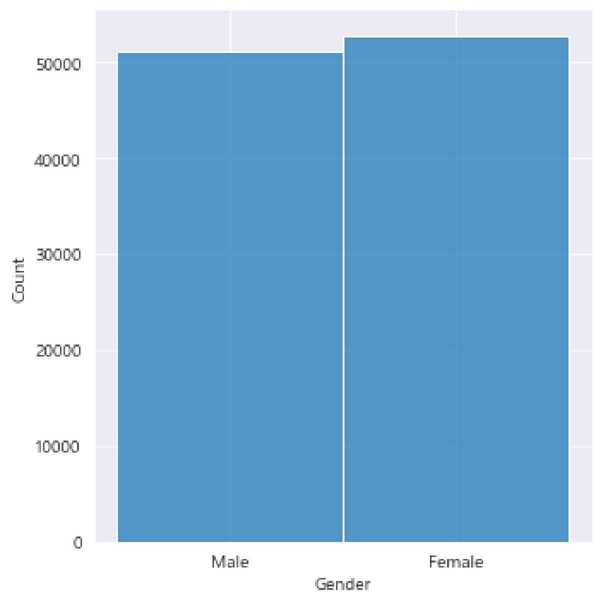


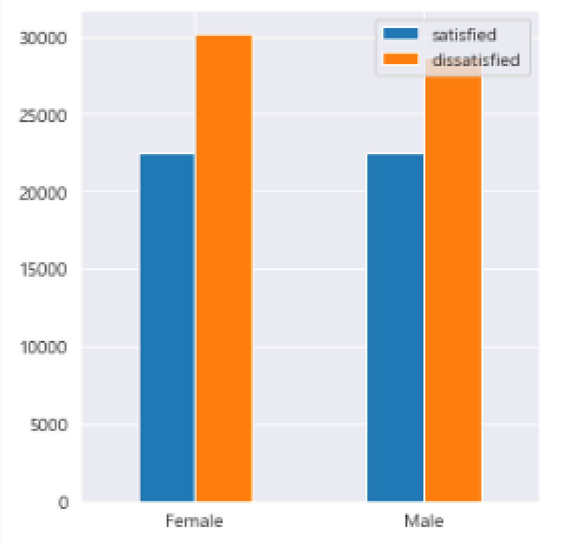
```
In [11]:   data.isnull().any()

Out[11]:  Unnamed: 0                           False
          id                                   False
          Gender                               False
          Customer Type                        False
          Age                                  False
          Type of Travel                       False
          Class                                False
          Flight Distance                      False
          Inflight wifi service                False
          Departure/Arrival time convenient    False
          Ease of Online booking               False
          Gate location                        False
          Food and drink                       False
          Online boarding                      False
          Seat comfort                         False
          Inflight entertainment               False
          On-board service                     False
          Leg room service                     False
          Baggage handling                     False
          Checkin service                      False
          Inflight service                     False
          Cleanliness                          False
          Departure Delay in Minutes           False
          Arrival Delay in Minutes              True
          satisfaction                         False
          dtype: bool
```

Arrival Delay in Minutes 변수에 결측값이 310개 존재함

# 3. 인구사회학적 요인

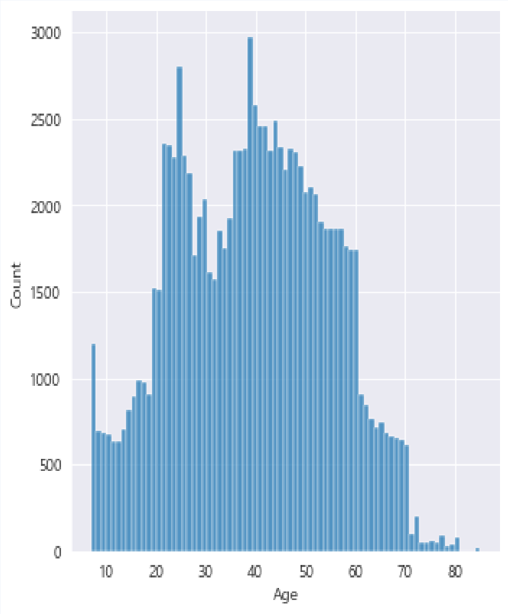
## 1. 성별

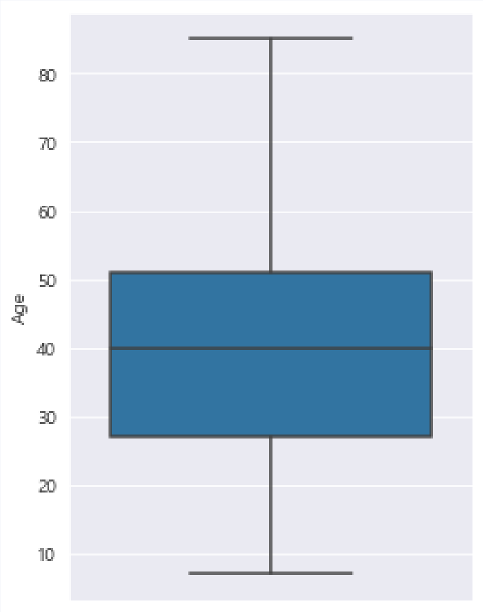


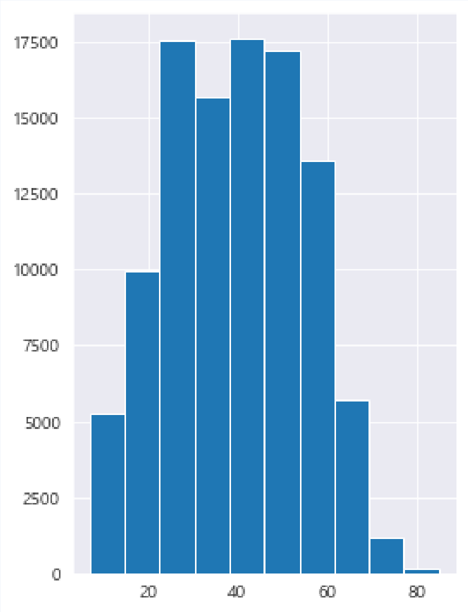


성별은 여성이 더 많으며,
주로 30~50대에서 많은 탑승객 분포를 보임.

## 2. 나이

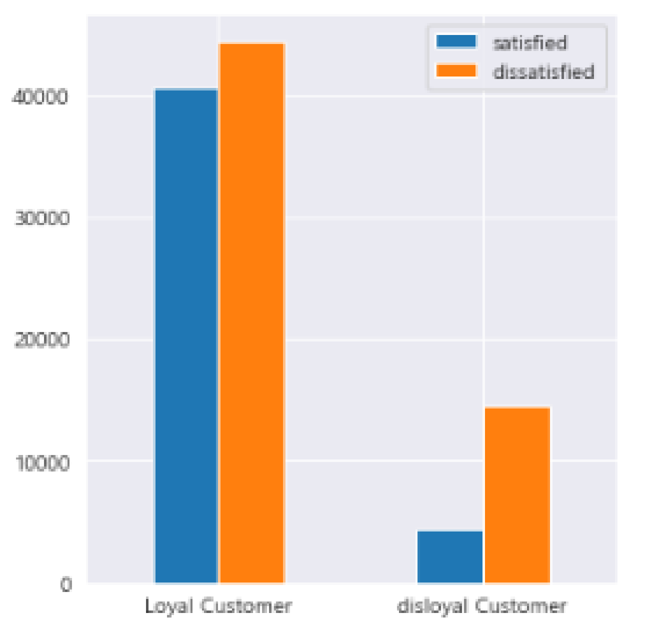




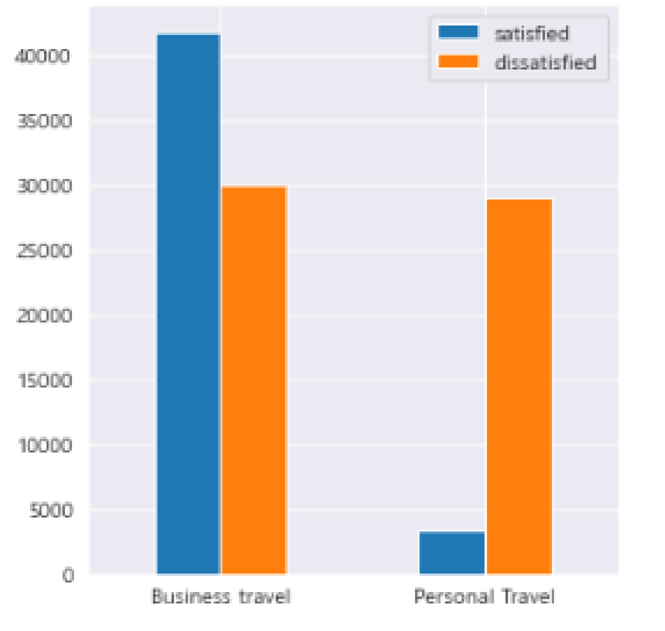


# 4. 승객 유형

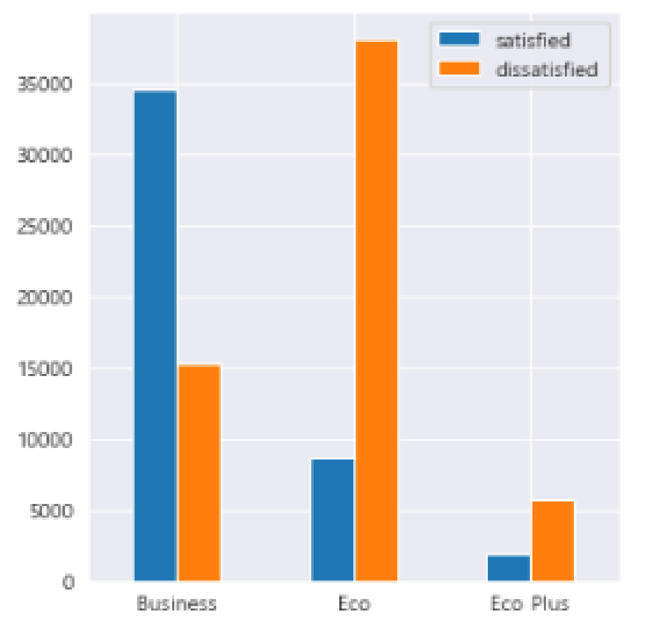
## 1. 고객 유형



## 2. 여행 유형



## 3. 좌석 유형

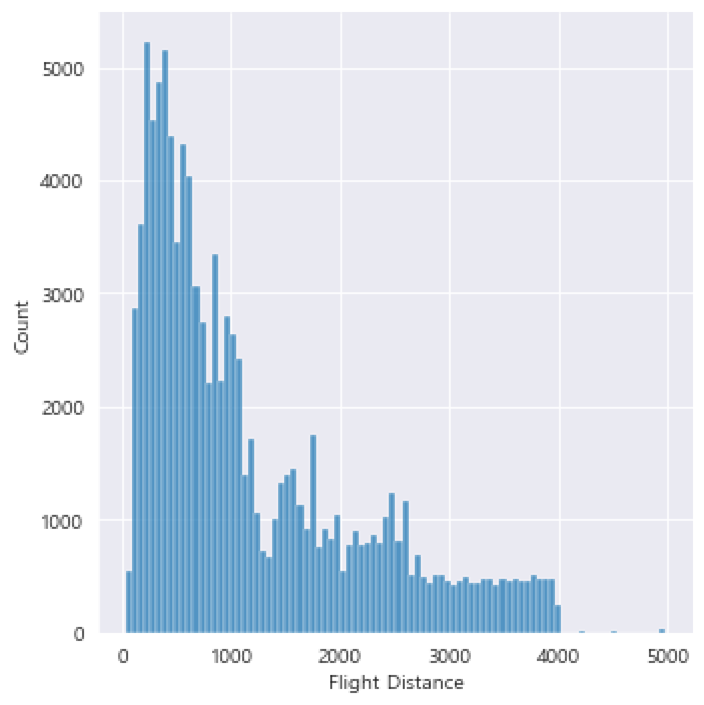


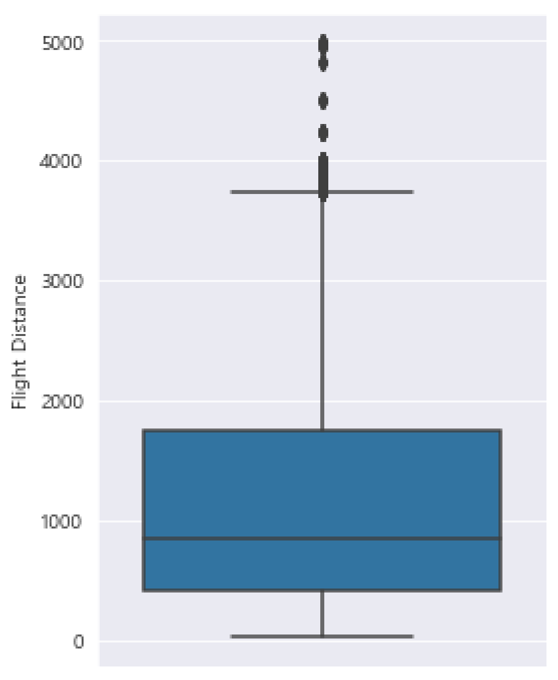
| 고객 유형 | | 여행 유형 | | 좌석 유형 | |
|---|---|---|---|---|---|
| Loyal Customer | 84923 | Business travel | 71655 | Business | 49665 |
| disloyal Customer | 18981 | Personal Travel | 32249 | Eco | 46745 |
| | | | | Eco Plus | 7494 |

비충성 고객, 개인목적일 경우, 좌석의 등급이 떨어질수록 불만족 비율이 큼.

# 4. 승객 유형

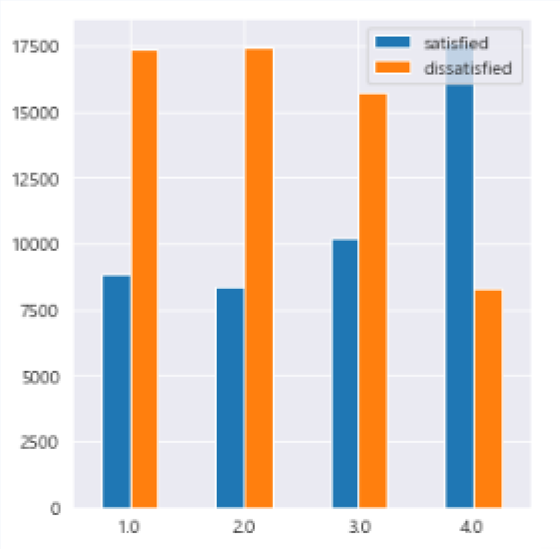
## 4. 비행거리





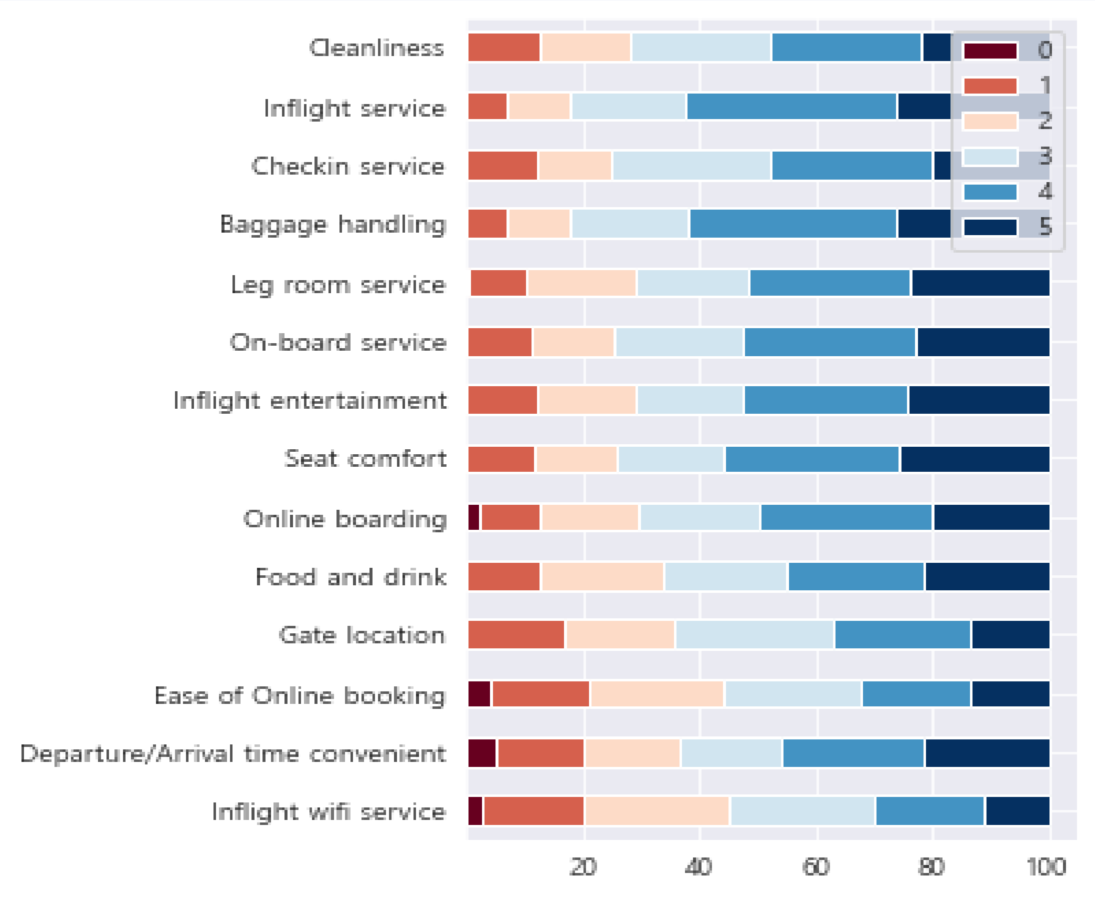
|  |  |
|---|---|
| count | 103904 |
| mean | 1189.4 |
| std | 997.1 |
| min | 31 |
| 25% | 414 |
| 50% | 843 |
| 75% | 1743 |
| max | 4983 |

오른쪽 방향으로 꼬리가 긴 분포. (skewed)



사분위수 범위로 나누어 처리해보았더니
비행거리가 길수록 만족하는 비율이 커짐.

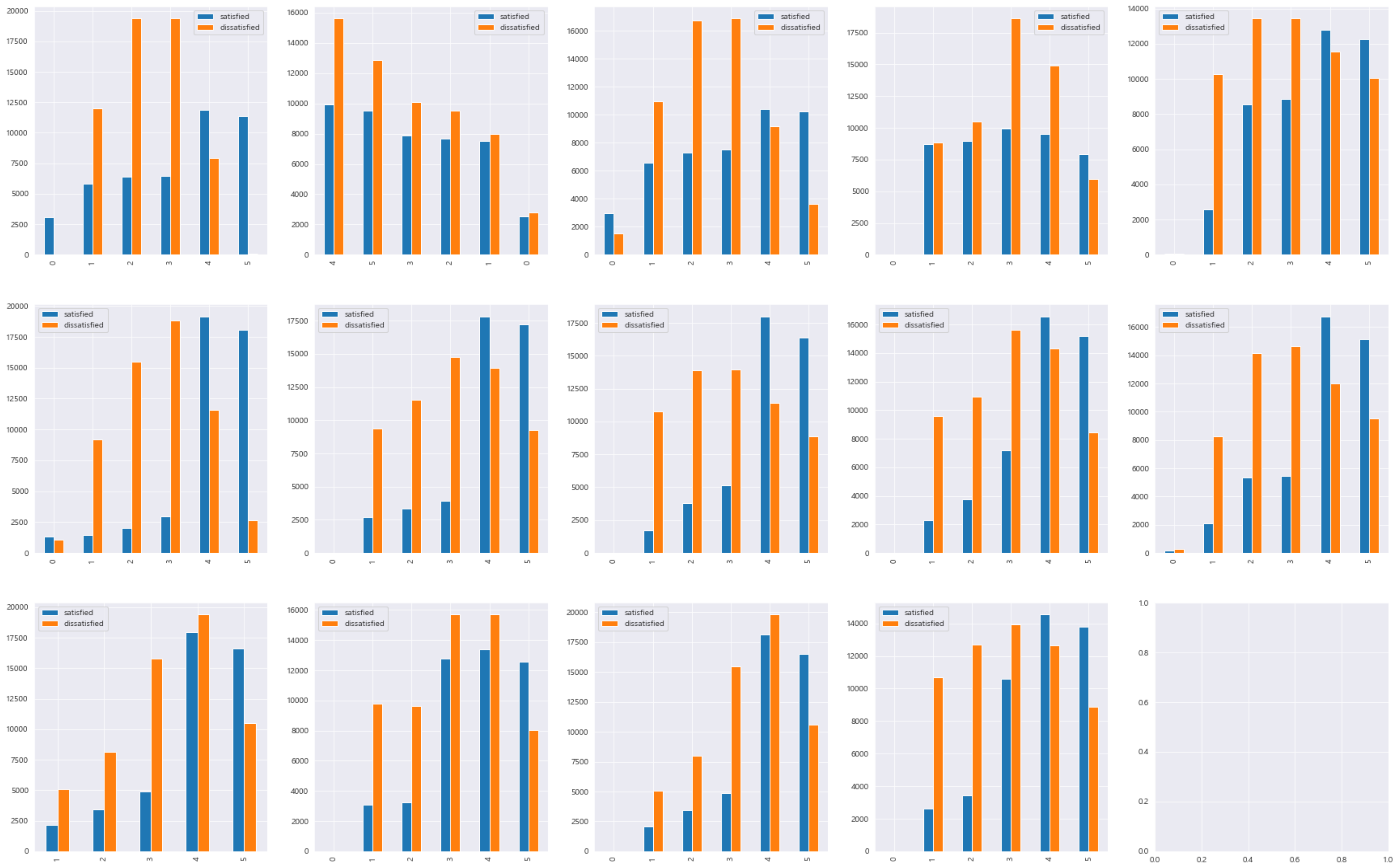
# 5. 개별 서비스 만족도 (5점 척도)

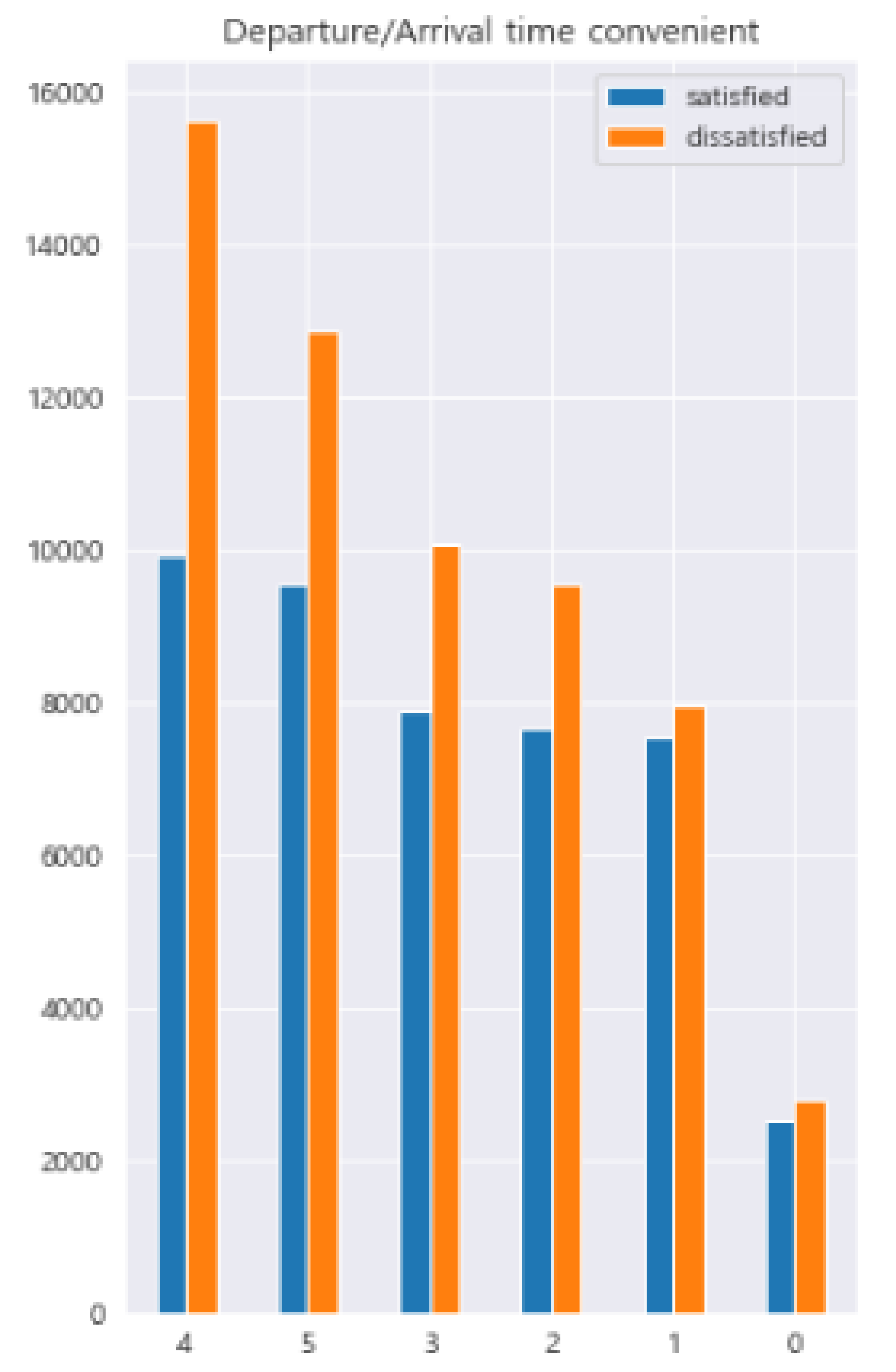


| Service | mean | std |
|---|---|---|
| Inflight wifi service | 2.73 | 1.33 |
| Departure/Arrival time convenient | 3.06 | 1.53 |
| Ease of Online booking | 2.76 | 1.40 |
| Gate location | 2.98 | 1.28 |
| Food and drink | 3.20 | 1.33 |
| Online boarding | 3.25 | 1.35 |
| Seat comfort | 3.44 | 1.32 |
| Inflight entertainment | 3.36 | 1.33 |
| On-board service | 3.38 | 1.29 |
| Leg room service | 3.35 | 1.32 |
| Baggage handling | 3.63 | 1.18 |
| Checkin service | 3.30 | 1.27 |
| Inflight service | 3.64 | 1.18 |
| Cleanliness | 3.29 | 1.31 |

개별항목 만족도가 높을수록 5점에 가까워지고, 낮을수록 1점에 가까워짐.
0점은 해당 없음 (Not Applicable)

크론바흐-알파값은 0.772로 일정한 수준의 내적일관성을 보임.

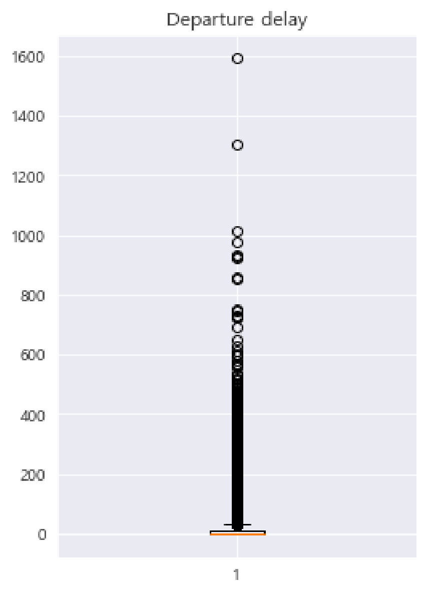

satisfied
dissatisfied
0
2500
5000
7500
10000
12500
15000
17500
20000
0
1
2
3
4
5
satisfied
dissatisfied
0
2000
4000
6000
8000
10000
12000
14000
16000
4
5
3
2
1
0
satisfied
dissatisfied
0
2000
4000
6000
8000
10000
12000
14000
16000
0
1
2
3
4
5
satisfied
dissatisfied
0
2500
5000
7500
10000
12500
15000
17500
0
1
2
3
4
5
satisfied
dissatisfied
0
2000
4000
6000
8000
10000
12000
14000
0
1
2
3
4
5
satisfied
dissatisfied
0
2500
5000
7500
10000
12500
15000
17500
20000
0
1
2
3
4
5
satisfied
dissatisfied
0
2500
5000
7500
10000
12500
15000
17500
0
1
2
3
4
5
satisfied
dissatisfied
0
2500
5000
7500
10000
12500
15000
17500
0
1
2
3
4
5
satisfied
dissatisfied
0
2000
4000
6000
8000
10000
12000
14000
16000
0
1
2
3
4
5
satisfied
dissatisfied
0
2000
4000
6000
8000
10000
12000
14000
16000
0
1
2
3
4
5
satisfied
dissatisfied
0
2500
5000
7500
10000
12500
15000
17500
20000
1
2
3
4
5
satisfied
dissatisfied
0
2000
4000
6000
8000
10000
12000
14000
16000
0
1
2
3
4
5
satisfied
dissatisfied
0
2500
5000
7500
10000
12500
15000
17500
20000
0
1
2
3
4
5
satisfied
dissatisfied
0
2000
4000
6000
8000
10000
12000
14000
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

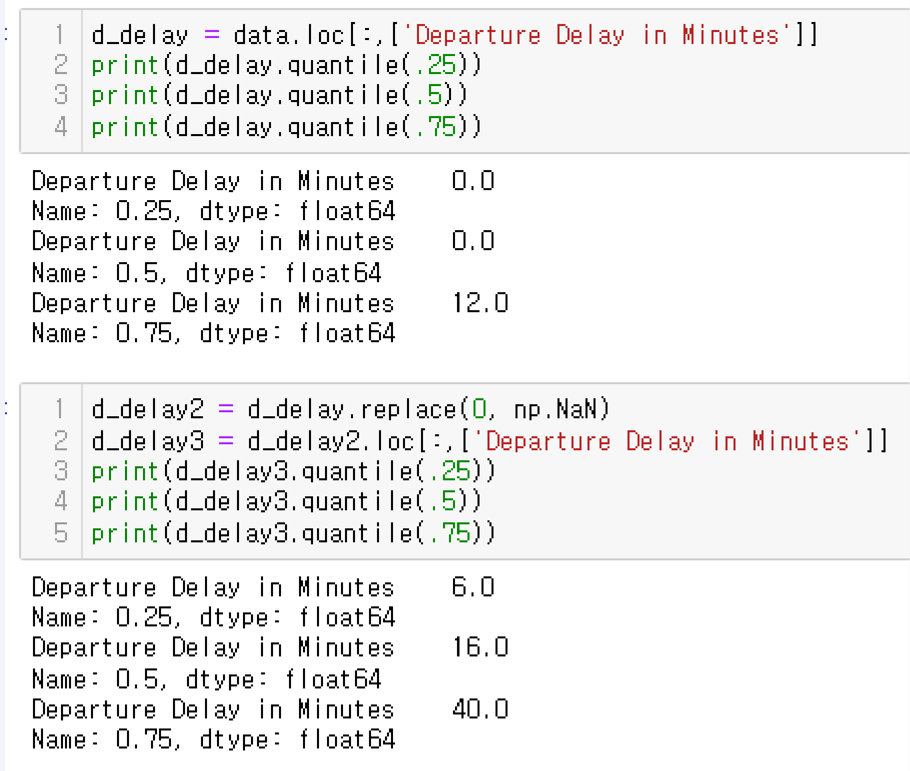


대부분 개별 서비스 만족도가 높을수록 최종 만족도가 높아지는 경향을 보이나,

유일하게 Departure/Arrival time convenient 만족도 변수는 최종 만족도와 반비례 관계를 보임.

# 6. 이/착륙 지연 시간



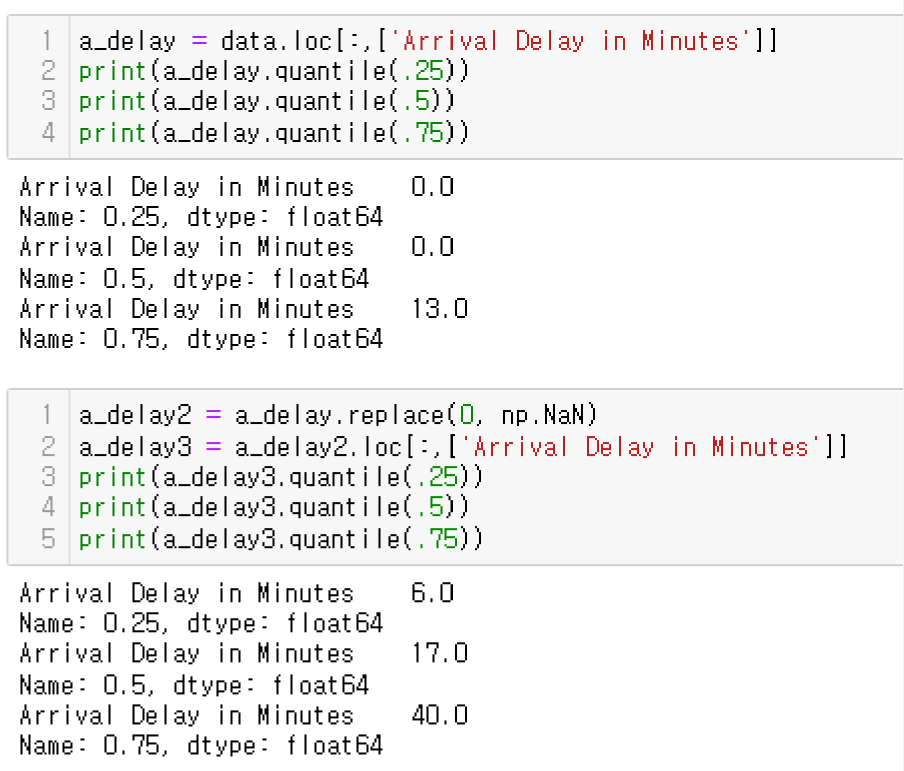


```
d_delay = data.loc[:,['Departure Delay in Minutes']]
print(d_delay.quantile(.25))
print(d_delay.quantile(.5))
print(d_delay.quantile(.75))
```

```
Departure Delay in Minutes    0.0
Name: 0.25, dtype: float64
Departure Delay in Minutes    0.0
Name: 0.5, dtype: float64
Departure Delay in Minutes    12.0
Name: 0.75, dtype: float64
```

```
d_delay2 = d_delay.replace(0, np.NaN)
d_delay3 = d_delay2.loc[:,['Departure Delay in Minutes']]
print(d_delay3.quantile(.25))
print(d_delay3.quantile(.5))
print(d_delay3.quantile(.75))
```

```
Departure Delay in Minutes    6.0
Name: 0.25, dtype: float64
Departure Delay in Minutes    16.0
Name: 0.5, dtype: float64
Departure Delay in Minutes    40.0
Name: 0.75, dtype: float64
```



```
a_delay = data.loc[:,['Arrival Delay in Minutes']]
print(a_delay.quantile(.25))
print(a_delay.quantile(.5))
print(a_delay.quantile(.75))
```

```
Arrival Delay in Minutes    0.0
Name: 0.25, dtype: float64
Arrival Delay in Minutes    0.0
Name: 0.5, dtype: float64
Arrival Delay in Minutes    13.0
Name: 0.75, dtype: float64
```

```
a_delay2 = a_delay.replace(0, np.NaN)
a_delay3 = a_delay2.loc[:,['Arrival Delay in Minutes']]
print(a_delay3.quantile(.25))
print(a_delay3.quantile(.5))
print(a_delay3.quantile(.75))
```

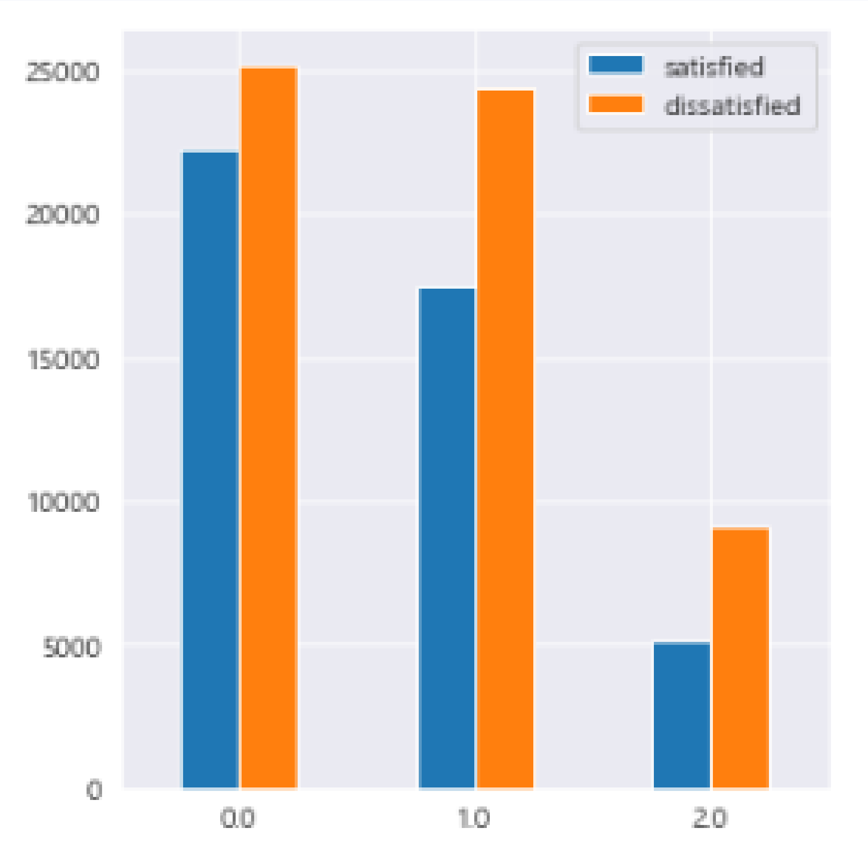
```
Arrival Delay in Minutes    6.0
Name: 0.25, dtype: float64
Arrival Delay in Minutes    17.0
Name: 0.5, dtype: float64
Arrival Delay in Minutes    40.0
Name: 0.75, dtype: float64
```

| | Departure Delay in Minutes | Arrival Delay in Minutes |
|---|---|---|
| mean | 14.81 | 15.17 |
| std | 38.23 | 38.69 |
| min | 0 | 0 |
| 25% | 0 | 0 |
| 50% | 0 | 0 |
| 75% | 12 | 13 |
| max | 1592 | 1584 |

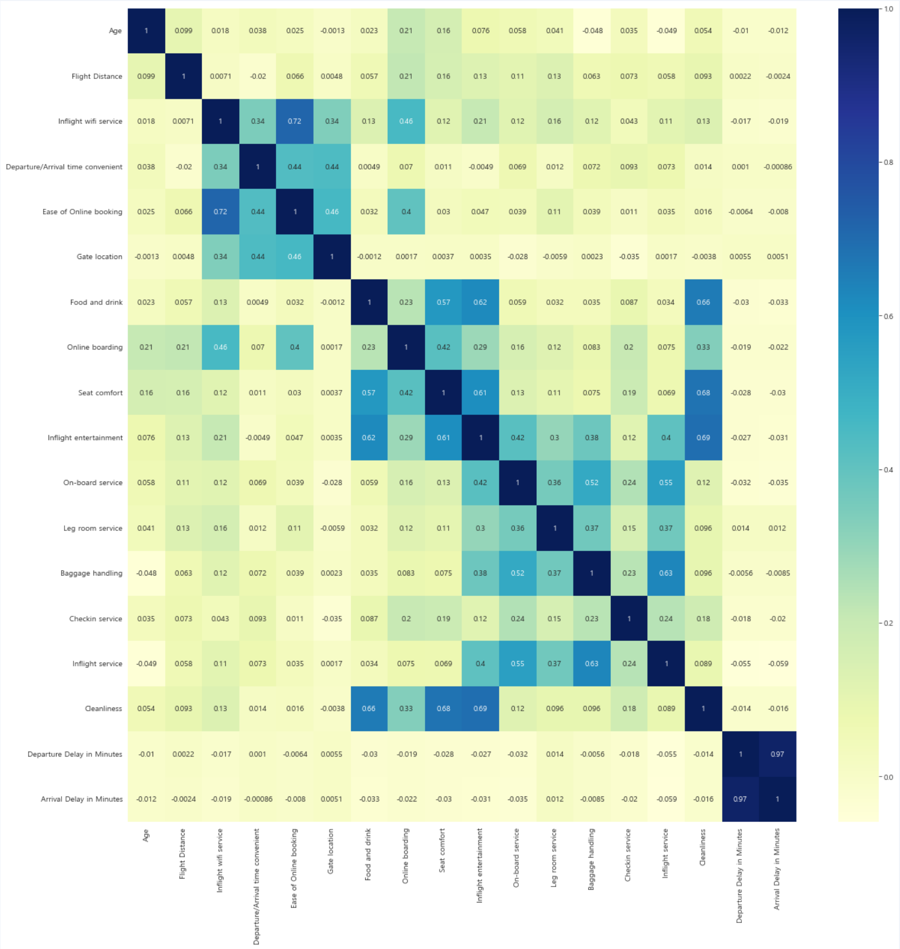
대부분의 경우는 지연시간이 없으며
지연시간이 존재하는 경우도 대다수는 1시간을 넘어가지 않음.

# 6. 이/착륙 지연 시간

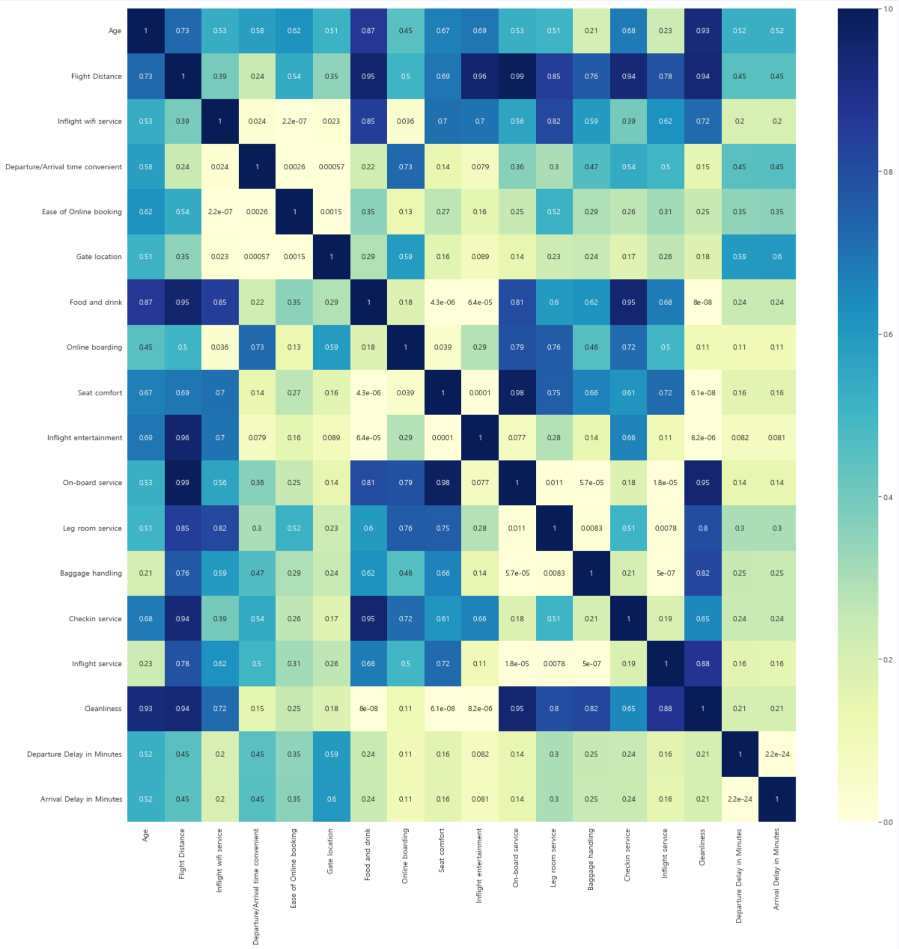
이륙 지연시간과 착륙 지연시간을 합쳐 새로운 파생변수를 생성하고
지연시간이 없는 경우, 지연시간이 60분 이하인 경우, 지연시간이 60분 이상인 경우, 총 3가지로 나누어보았음.



# 7. 상관계수 히트맵



Coefficient



P-value