기초 이론부터 실무 실습까지

머신 러닝 익히기

programmers

Part 04. 선형 회귀와 선형 분류

정 정 민

programmers

Chapter 11. 선형 분류 실습

1. 비행 경험 만족도 데이터
2. EDA, 탐색적 데이터 분석
3. 데이터 전처리
4. 모델 구축 및 결과 확인

비행 경험 만족도 데이터

비행 경험 만족도 데이터 (Airlines Customer satisfaction)

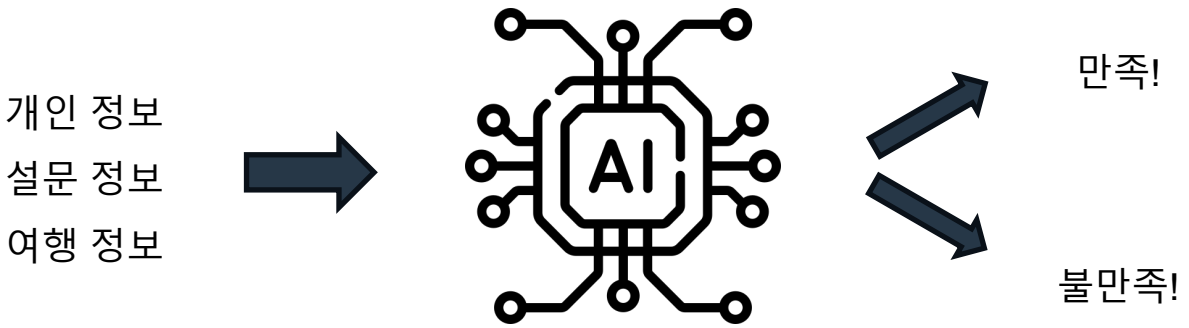
- 이번 실습에서 사용할 데이터로 Kaggle의 공개 데이터 (링크)
  - 다운로드 받아주세요!

- 항공사 서비스에 대한 고객 만족도 관련 데이터
- 아래의 변수(총 23개)를 포함
  - 만족도
  - 나이, 비행 거리, 출발 지연 시간, 도착 지연 시간
  - 좌석 편안함, 출발/도착 시간 편리함, 음식 및 음료, 탑승구 위치, 기내 와이파이 서비스, 기내 엔터테인먼트, 온라인 지원, 온라인 예약의 용이성, 기내 서비스, 다리 공간, 수하물 처리, 체크인 서비스, 청결도, 온라인 탑승
  - 성별, 고객 유형, 여행 유형, 등급

문제 정의

- 풀어야 하는 문제
  - 주어진 탑승객의 개인 및 여행 경험 정보를 바탕으로 전반적인 비행의 만족도를 예측
    독립 변수 종속 변수

- 머신 러닝 모델의 입, 출력 정의
  - 입력 : 앞선 독립 변수들
    - 개인 정보
    - 설문 정보
    - 여행 정보
  - 출력 : 비행의 만족도 (종속 변수)



EDA, 탐색적 데이터 분석

programmers

강의 실습 코드 : 링크

기본 정보

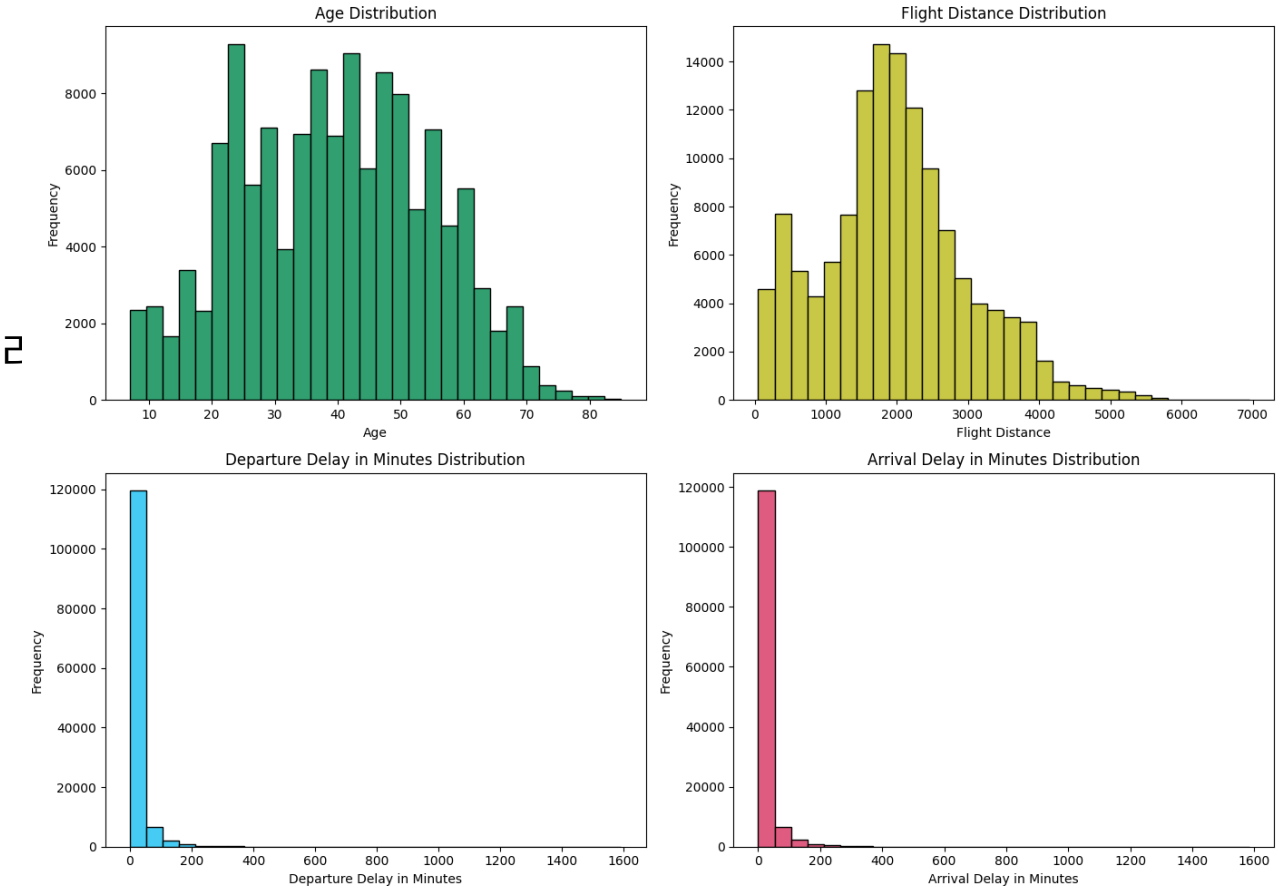
- 전체 데이터셋 크기
  - 총 129,880개의 개별 데이터
  - 총 23개 특성
- 데이터 타입
  - 수치형 (Numerical)
  - 서수형 (Ordinal)
    - 순서나 등급을 나타냄
    - 순서는 중요하지만 그 차이는 균일하지 않음
    - 설문 조사, 학점, 통증 수준 등
  - 범주형 (Categorical)
- 누락 : Arrival Delay in Minutes

```
RangeIndex: 129880 entries, 0 to 129879
Data columns (total 23 columns):
 #   Column                            Non-Null Count   Dtype
---  ------                            --------------   -----
 0   satisfaction                      129880 non-null  object
 1   Gender                            129880 non-null  object
 2   Customer Type                     129880 non-null  object
 3   Age                               129880 non-null  int64
 4   Type of Travel                    129880 non-null  object
 5   Class                             129880 non-null  object
 6   Flight Distance                   129880 non-null  int64
 7   Seat comfort                      129880 non-null  int64
 8   Departure/Arrival time convenient 129880 non-null  int64
 9   Food and drink                    129880 non-null  int64
 10  Gate location                     129880 non-null  int64
 11  Inflight wifi service             129880 non-null  int64
 12  Inflight entertainment            129880 non-null  int64
 13  Online support                    129880 non-null  int64
 14  Ease of Online booking            129880 non-null  int64
 15  On-board service                  129880 non-null  int64
 16  Leg room service                  129880 non-null  int64
 17  Baggage handling                  129880 non-null  int64
 18  Checkin service                   129880 non-null  int64
 19  Cleanliness                       129880 non-null  int64
 20  Online boarding                   129880 non-null  int64
 21  Departure Delay in Minutes        129880 non-null  int64
 22  Arrival Delay in Minutes          129487 non-null  float64
dtypes: float64(1), int64(17), object(5)
memory usage: 22.8+ MB
```

시각화 – 수치형 데이터

- 그래프 시각화를 통해
  - 전체 데이터의 분포 확인
  - 만족도 결과에 따른 분포 확인

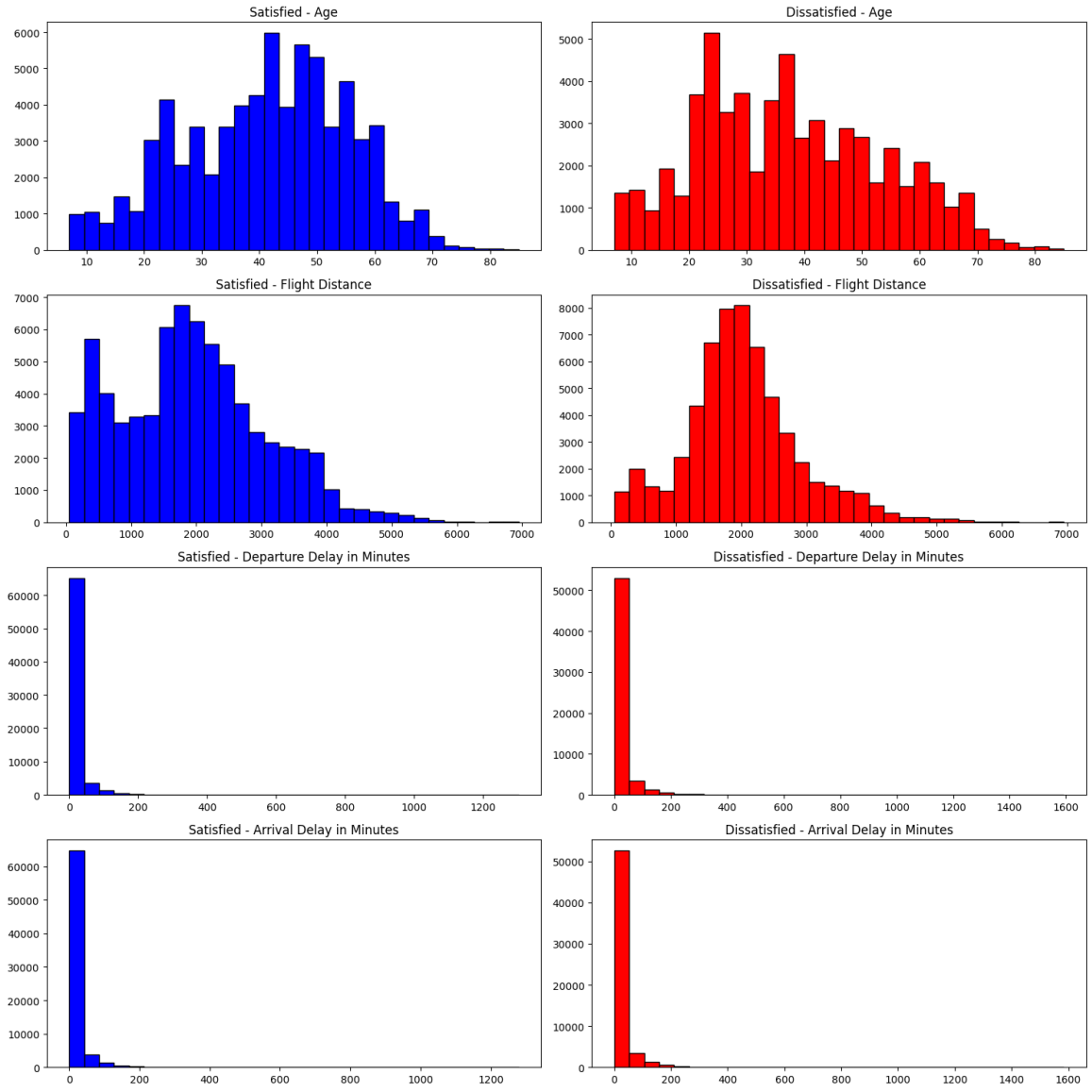
- 출발 및 도착 지연 시간에는 outlier가 보임
- 하지만 이 결과가 만족도에 영향을 줄 수 있음을 고려



시각화 – 수치형 데이터

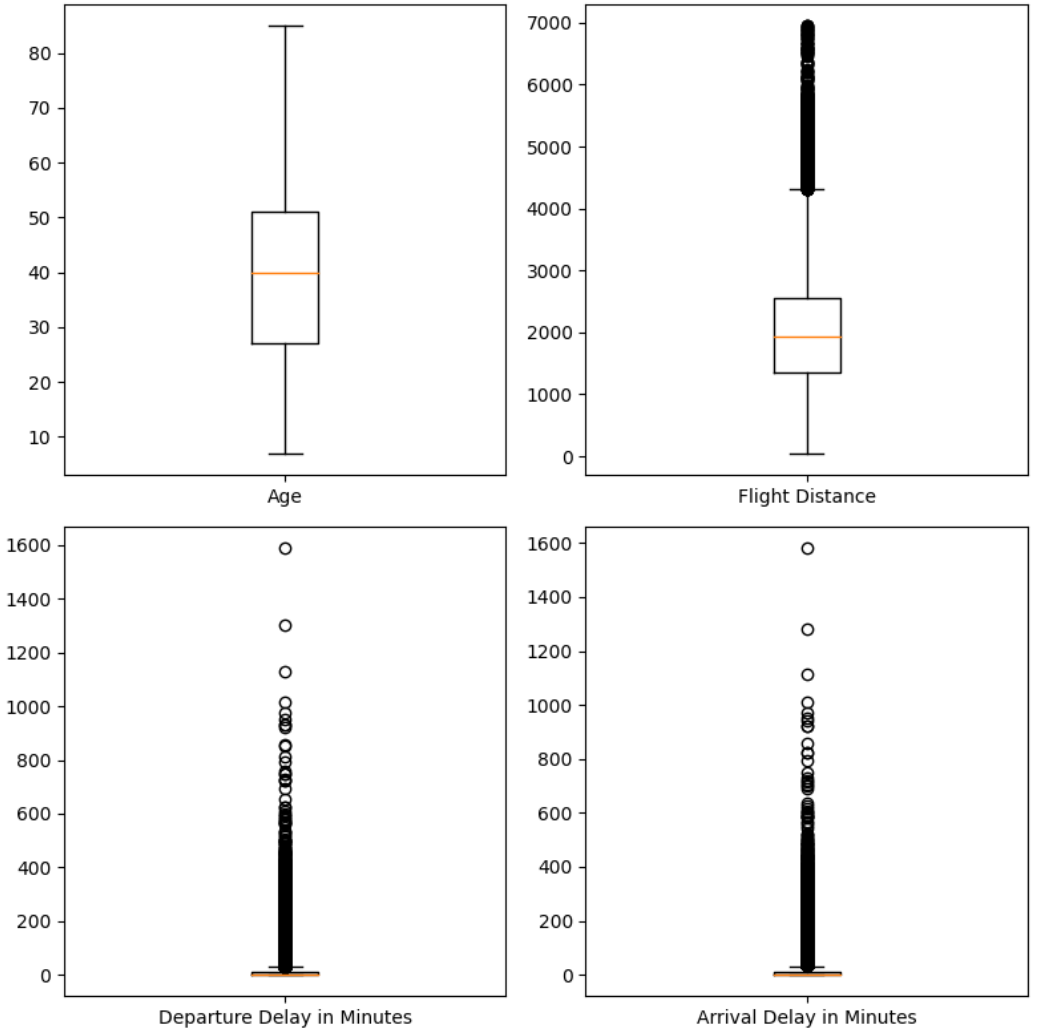
- 그래프 시각화를 통해
  - 전체 데이터의 분포 확인
  - 만족도 결과에 따른 분포 확인

- 출발 및 도착 지연 시간에는 outlier가 보임
- 하지만 이 결과가 만족도에 영향을 줄 수 있음을 고려



아웃라이어 – 수치형 데이터

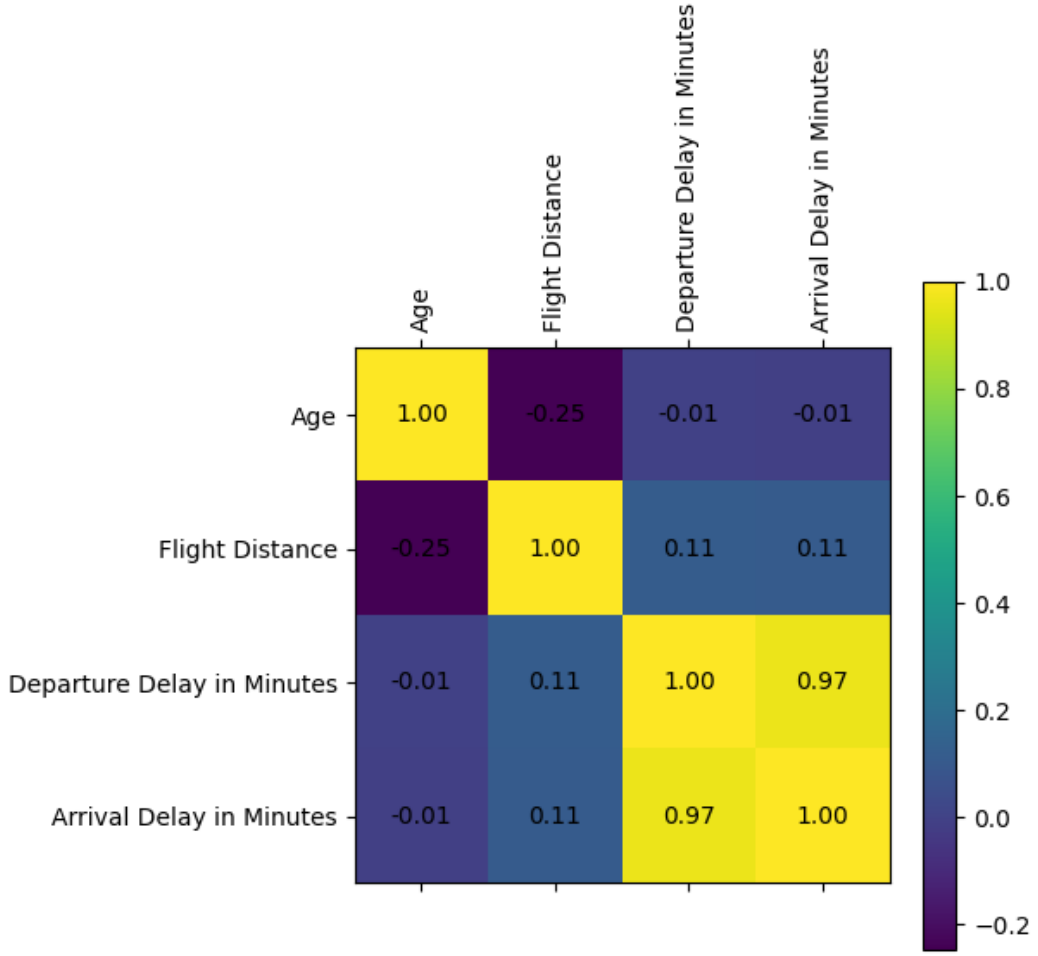
- Age
  - 전체 모형이 정규 분포 모양
  - 균일한 분포이며 아웃라이어는 안보임
- Flight Distance
  - 비교적 균일, 약간의 아웃라이어
  - 하지만 충분히 있을 수 있는 정도
- Delay time
  - 극단적인 값이 존재



상관관계 – 수치형 데이터

- 출발 시간이 늦으면 도착 시간도 늦음
- 두 변수 사이에 매우 큰 상관관계 예상

- 선형 모델에는 부정적 영향을 미칠 수 있음을 인지

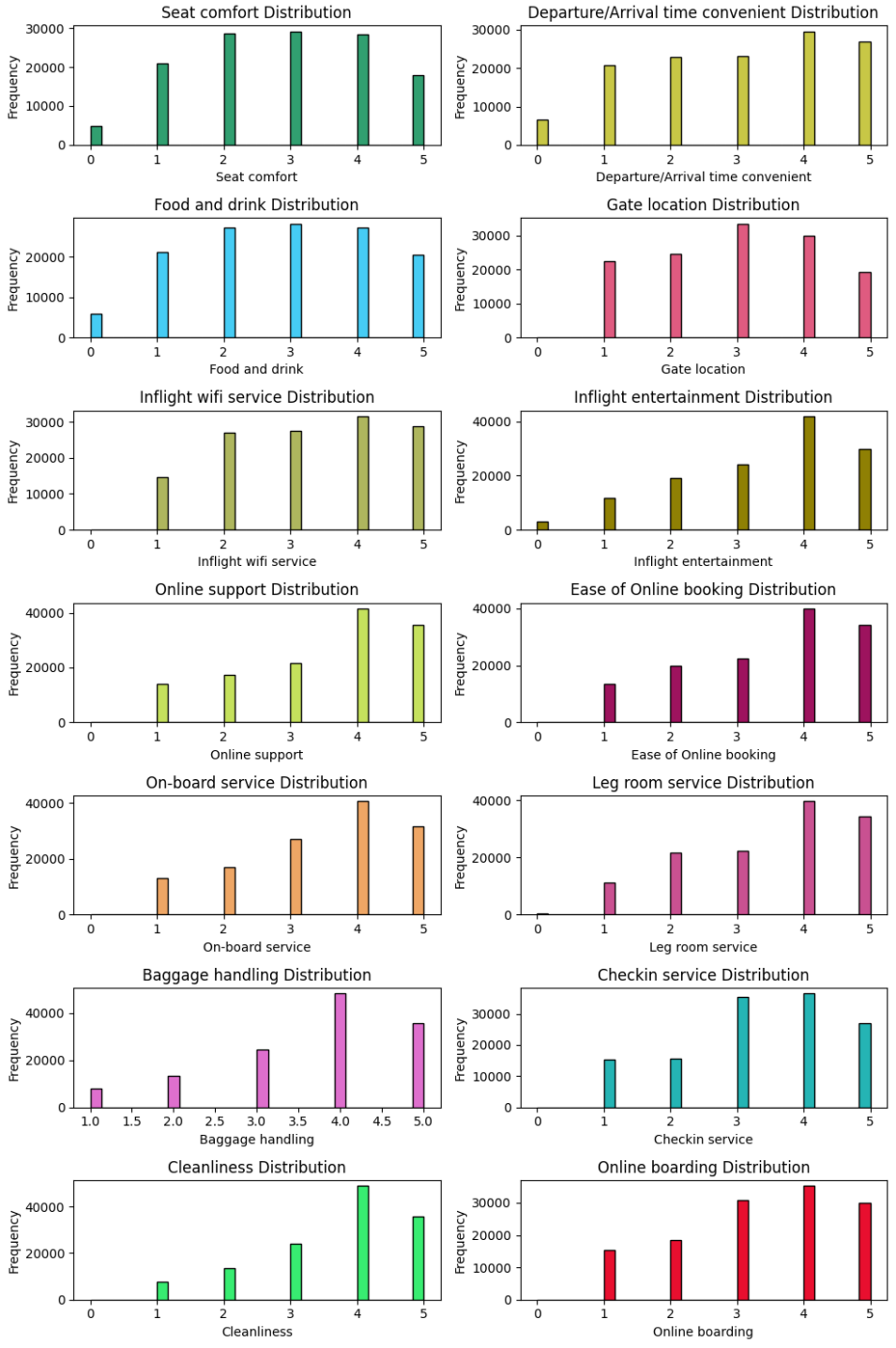


시각화 – 서수형 데이터

- 0~5점에 해당하는 설문 평가 데이터의 시각화
- 극단적으로 답변이 치우진 문항은 없어 보임

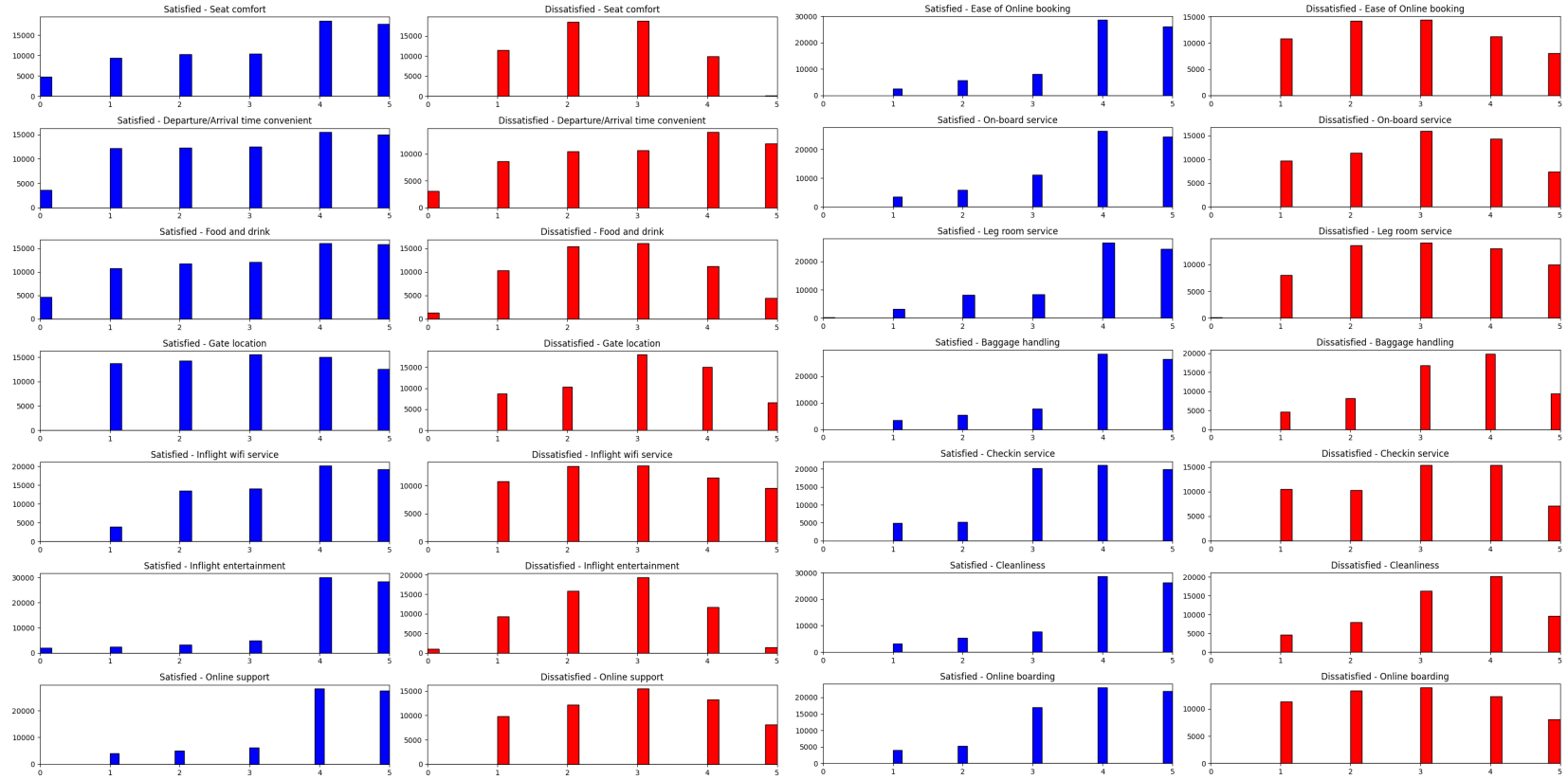
- 따라서 **모델 개발 입장**에서
  - 특별히 주의할 변수는 보이지 않음
  - 대신 각 변수 마다 보이는 분포는 상이함
    - 상위점에 몰림 현상
    - 중간 점수에 몰림 현상

- **서비스 개선의 입장**에서
  - 잘 하고 있는 것과
  - 개선이 필요한 부분 (Food & Drink, Seat Comfort 등)



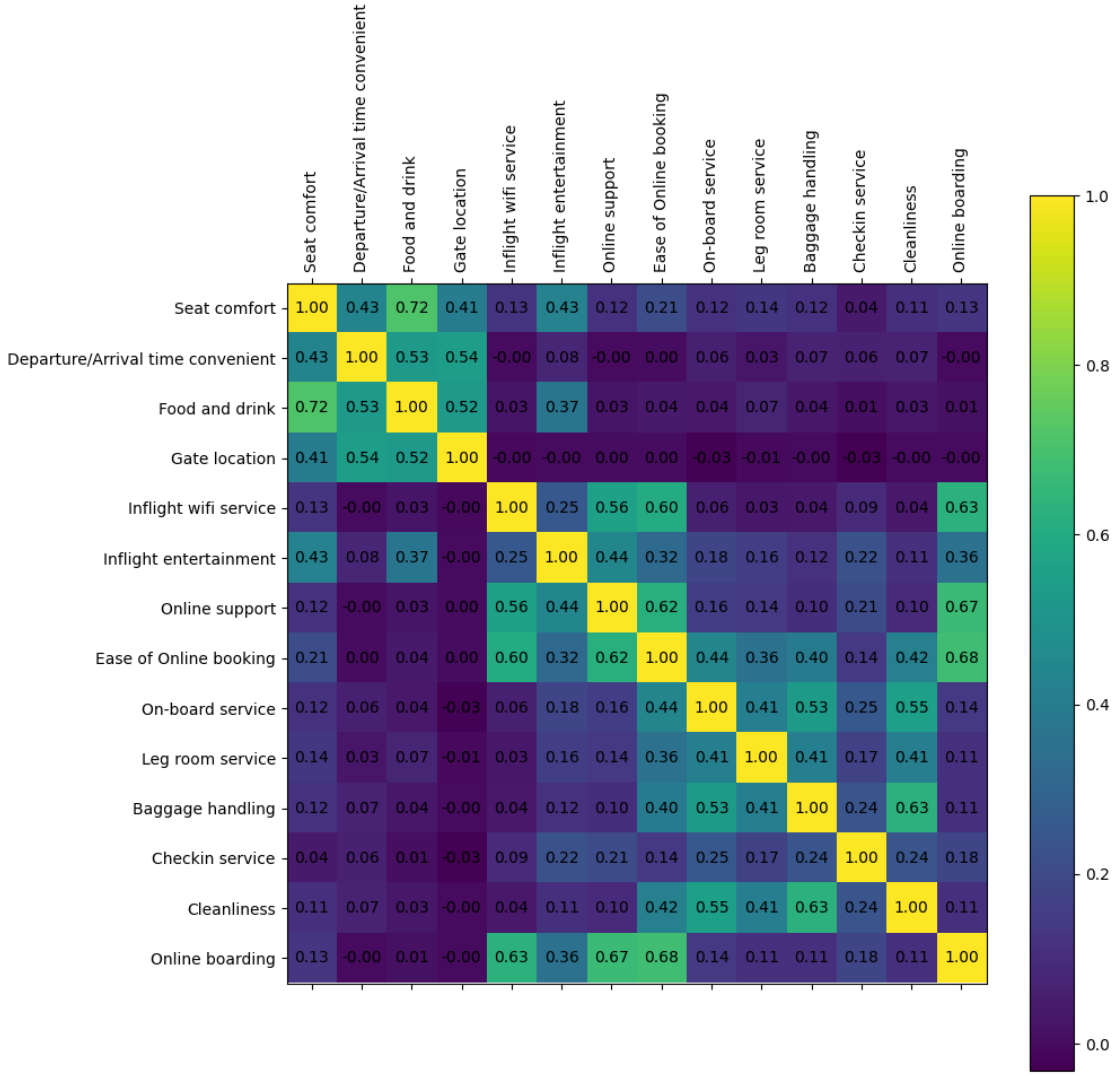
시각화 – 서수형 데이터

- 만족도 결과에 따른 각 서수형 변수 시각화
- 만족 / 불만족에 따른 분포 차이가 존재하는 경우(seat comfort, food and drink 등)도 있고
  그렇지 않은 경우(depart/arrival time comfort)도 존재
- 종속 변수의 결과로 나누어 확인할 경우,
  분포의 차이가 보이는 변수의 경우 분석에 큰 영향을 미칠 가능성 있음



상관관계 – 서수형 데이터

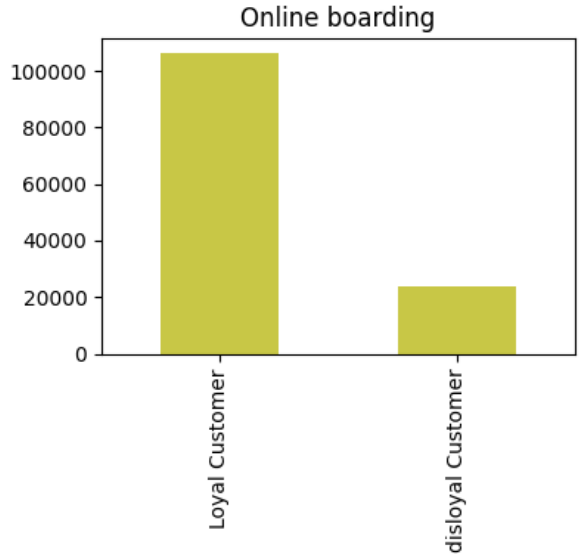
- 특정 문항 사이에 높은 상관관계가 보임
- 하지만 다른 변수를 대체할 만큼 상관 관계 값이 크지는 않아보임

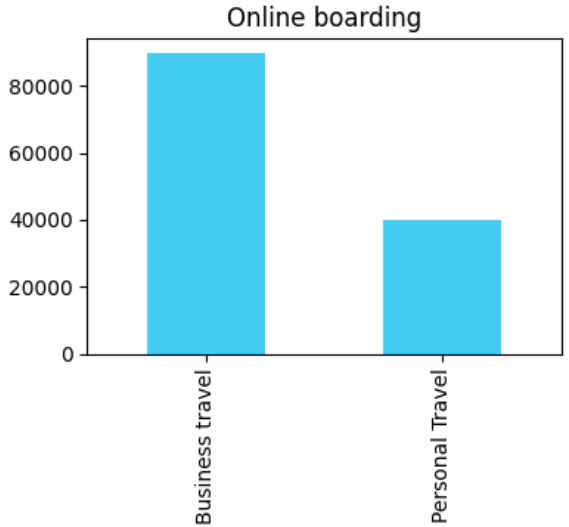


시각화 – 범주형 데이터

- 전체 데이터에 대해 특정 범주의 빈도수를 확인

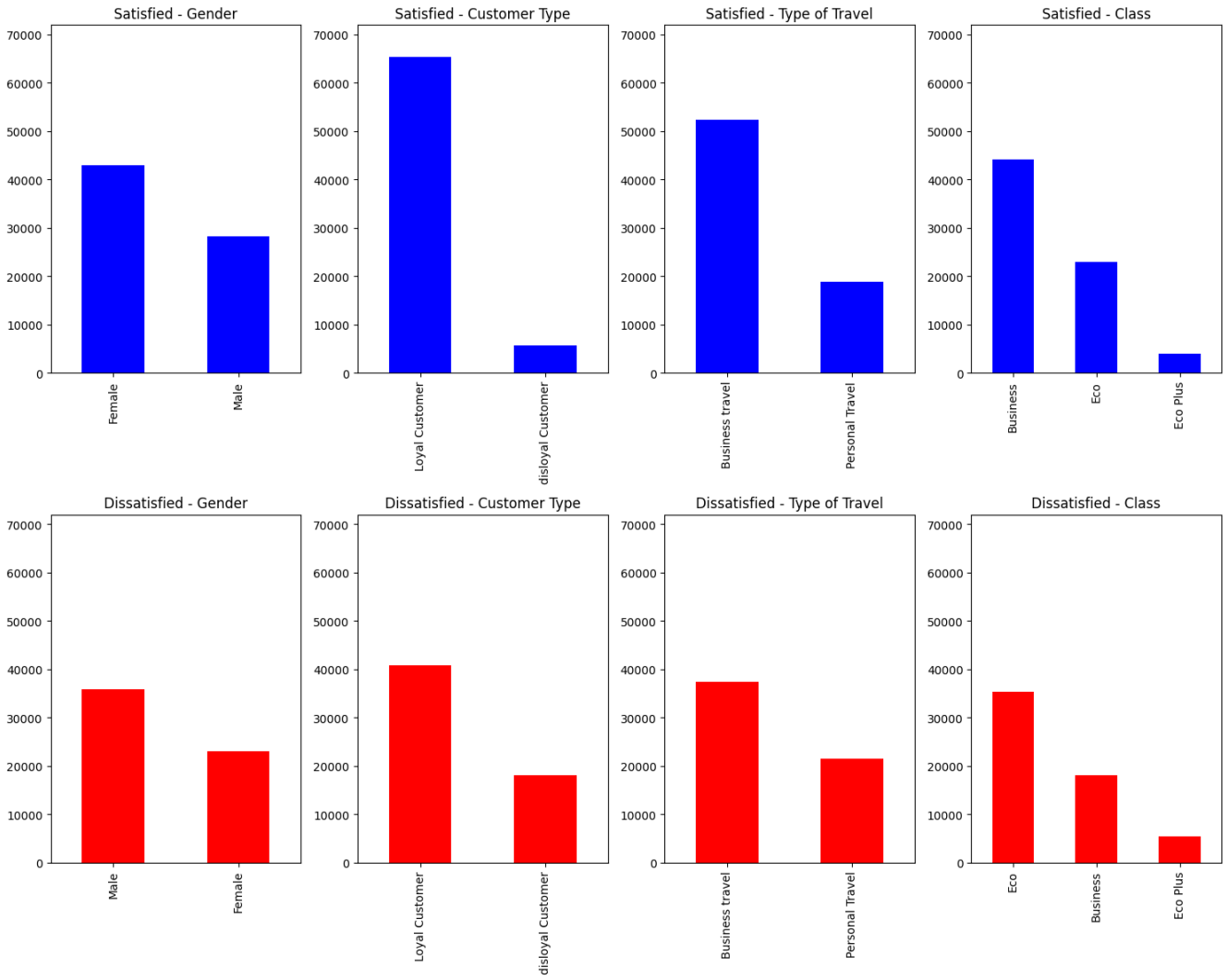
- 성별 : 응답 승객에서 성별 비율은 큰 차이가 없음
- 나머지 범주 데이터
  - 설문을 포함한 만족도 조사에 있어 응답자 차이가 보임
  - VIP / 일반 고객에서도 차이가 보임
  - 비지니스 고객이 더 응답을 많이 했고
  - Eco plus가 적음, 다만 이는 절대적인 수가 적을 수 있음





시각화 – 범주형 데이터

- 만족과 불만족 데이터 사이의 범주형 데이터 분포 확인
- 이 분포의 차이가 강한 범주일수록 분석에 큰 영향을 미칠 수 있음



데이터 전처리

강의 실습 코드 : 링크

제외 데이터 판단 및 제거!

- 누락 데이터를 포함한 데이터 포인트는 제거
- 또한, EDA 과정에서 판단한 제외 데이터를 제거
  - 이 과정은 정답은 없음
  - 이번 실습에서는 delay 시간을 기준으로 clipping 진행
- 결과적으로 129,149 데이터를 사용

```
airplane_cleaned = airplane.dropna() # na값 제거
time_limit = 300 # 지연 시간 5시간 이상은 제거
airplane_cleaned = airplane_cleaned[(airplane_cleaned['Arrival Delay in Minutes'] < time_limit) &
                                    (airplane_cleaned['Departure Delay in Minutes'] < time_limit)]
```

카테고리형 변수 인코딩

- 선형 회귀 실습 과정과 비슷
- 타겟하는 범주형 데이터로는
  - (종속변수) '만족도'
  - (독립변수) '성별', '고객 유형', '여행 유형', '클래스'
- 이 과정에서 1개의 변수가 추가됨

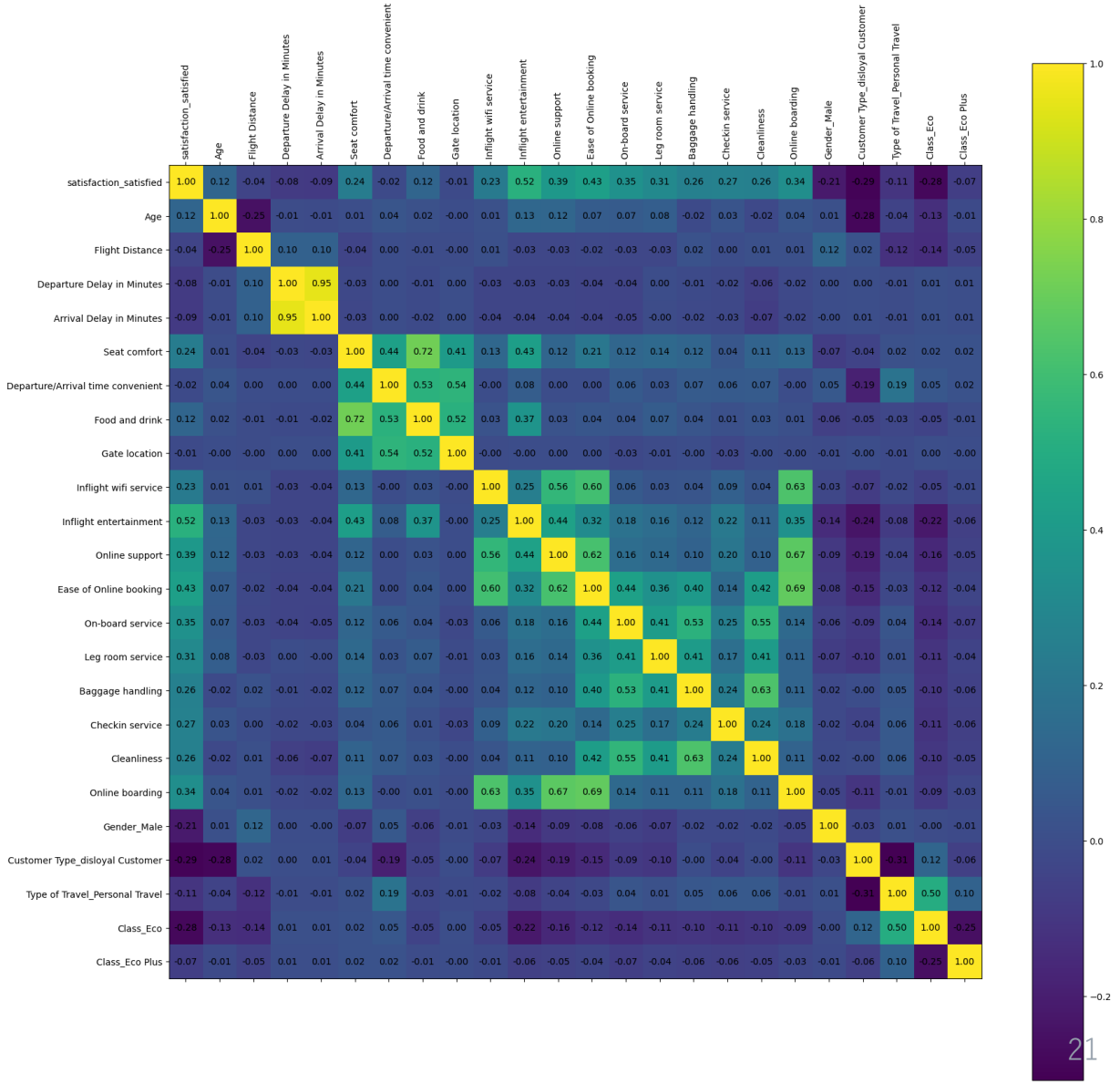
```
y_column = ['satisfaction']
category_columns = ['Gender', 'Customer Type', 'Type of Travel', 'Class']

airplane_cate_encoded = pd.get_dummies(airplane_cleaned[category_columns],
                                       drop_first=True)
airplane_target_encoded = pd.get_dummies(airplane_cleaned[y_column],
                                         drop_first=True)
```

일부 특성만 사용

- 종속 변수를 제외한 22개 변수 중
- 특정 변수와 큰 상관관계가 있는 변수도 존재
  - 선형 모델 가정에 문제를 일으킬 수 있음

- 해석력과 일반화 향상을 위해
  23개 변수 중 15개 변수만 취해서 학습 진행
- → 종속 변수와 여러 독립 변수 사이의 상관관계를 활용

- 상위 15개 변수를 취함



데이터 분할

- 80:20의 비율로 학습 및 평가 데이터로 분할

- 학습 데이터의 양 : 103,319개
- 평가 데이터의 양 : 25,830개

- 전체 특성 수 : 15개

```
from sklearn.model_selection import train_test_split

X = data.drop(y_column, axis=1)
y = data[y_column]
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.2,
                                                    random_state=42)
```

모델 구축 및 결과 확인

강의 실습 코드 : 링크

선형 분류 모델 학습 진행

- LogisticRegression() 객체를 생성
- Bias 항은 생성하지 자동 추가되므로 생성 x

- 총 15개 weight 학습
- + bias 까지 학습

```
from sklearn.linear_model import LogisticRegression

logistic_reg = LogisticRegression()
logistic_reg.fit(X_train, y_train)

coefficients = logistic_reg.coef_
intercept = logistic_reg.intercept_

##################### 학습된 파라미터 값 #####################
[[ 0.73145501  0.3166282   0.09574604  0.27898618  0.11109389  0.21935144
  -1.33676471 -0.93629662  0.25075678  0.06480377  0.04234803  0.25546098
  -0.08643957 -0.99928686 -0.25919007]]
##################### 학습된 절편 값 #####################
[-5.48451267]
```

평가 진행 (정확도, Accuracy)

- 예측한 결과가 실제 결과의 일치 혹은 불일치를 기반으로 정확도를 구할 수 있음
  - 전체 데이터로 일치 데이터의 수를 나눠줌

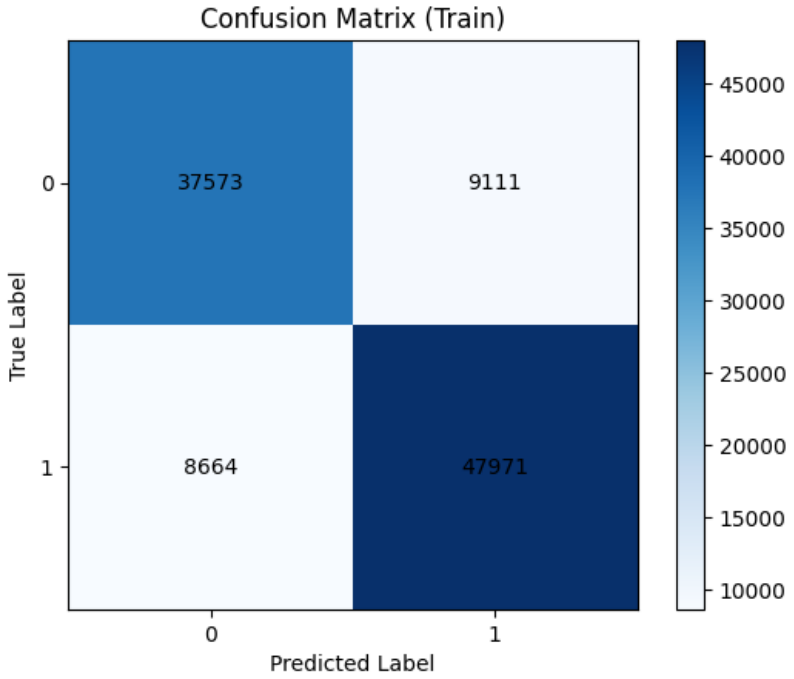
- 학습 데이터와 평가 데이터를 기준으로 평가 진행

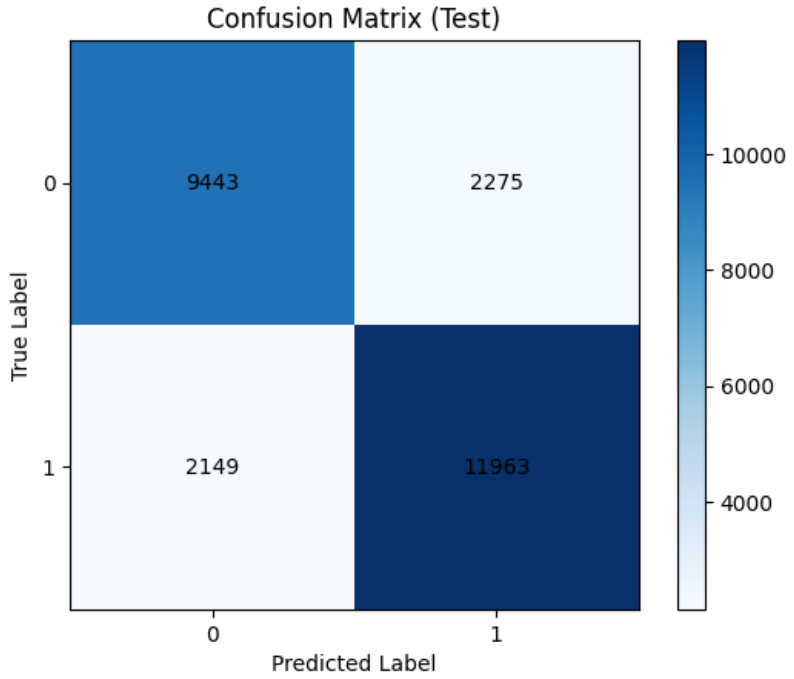
```python
from sklearn.metrics import accuracy_score

y_train_pred = logistic_reg.predict(X_train)
y_test_pred = logistic_reg.predict(X_test)

acc_train = accuracy_score(y_train, y_train_pred)
acc_test = accuracy_score(y_test, y_test_pred)

print('학습 데이터를 이용한 Acc 값 :', acc_train)
print('평가 데이터를 이용한 Acc 값 :', acc_test)

# 학습 데이터를 이용한 Acc 값 : 0.8279600073558591
# 평가 데이터를 이용한 Acc 값 : 0.8287262872628727
```

평가 진행 (Confusion Matrix)

- 예측 결과와 실제 결과의 일치 및 불일치를 기준으로 하지 않고
- 예측 그리고 실제 결과 “**값**”을 기준으로 표를 생성할 수 있음
  - 이를 Confusion Matrix 라고 함

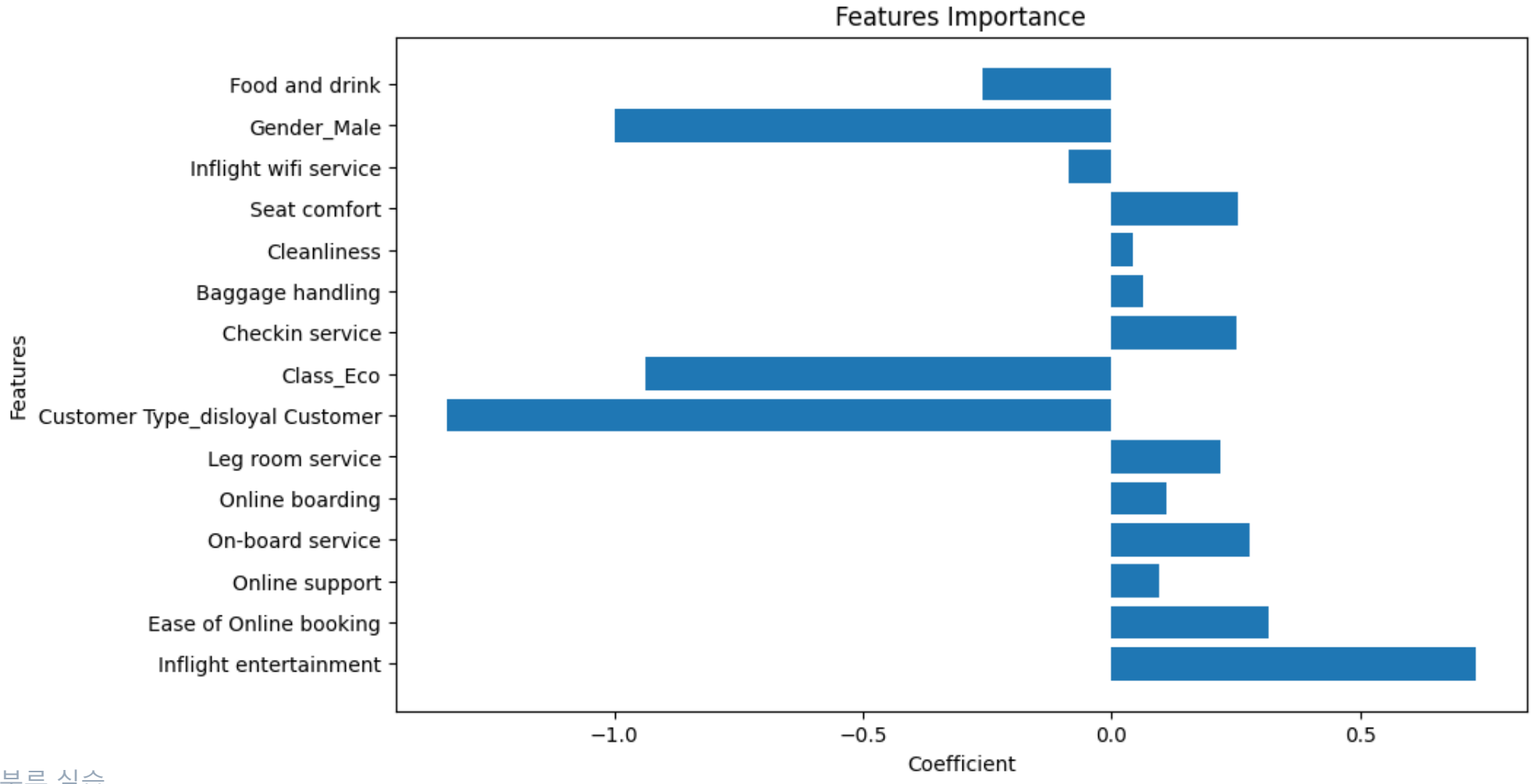
- 좌상 → 우하 방향 대각선 위치의 값이 클 수록 좋은 결과





결과 해석 – 변수의 중요도

- 파라미터를 시각화로 사용한 변수들의 중요도 확인



[숙제] 더 해보기!

- 변수 엔지니어링(Feature Engineering)을 통한 새로운 모델 학습
  - 모델의 입력에 해당하는 변수를 새롭게 정의해 다른 모델을 학습

- 아래 경우에 해당하는 학습을 진행해보고 각각의 결과를 비교해보세요.
  - 23개 모든 변수 전부 활용
  - 임의로 사용할 변수를 선택
  - EDA 분석을 통해
    - 제일 의미가 클 것 같다고 생각되는 변수 하나만 선택
    - 상위 K개만 선택해서 사용
  - 등등

E.O.D