항공사 서비스 만족도 데이터 기반 고객 세그멘테이션 및 서비스 개선 방안 도출 EDA

데브코스 데이터 분석 4기 1차 프로젝트 보고서

### 1. 서론

4차산업 혁명의 발전에 따라 매 순간순간 방대한 양의 데이터가 생성되고 있다. 이러한 빅데이터를 수집, 저장, 가공, 분석하는 기술의 필요성이 커지고 있으며, 이에 따라 데이터 분석 기술의 중요성이 점점 더 강조되고 있다. 데이터 분석은 유용한 정보를 발굴하고, 의사 결정에 도움을 주도록 데이터를 정리, 변환, 모델링하는 과정으로, 빅데이터를 분석하고 예측 모델을 도출하는 일련의 과정에서 필수적이다. 이 보고서에서는 데이터 분석 방법론 중 하나인 탐색적 데이터 분석(EDA)를 수행하여 의미 있는 결과를 도출하고, 데이터 간의 관계를 파악함으로써 데이터 분석 역량을 키우고자 한다

#### 데이터 선택 과정

데이터 분석과 머신러닝 학습 및 대회 플랫폼인 Kaggle에서 팀원 별로 각자 흥미와 관심 분야에 따라서 EDA를 수행하고 싶은 데이터 3가지를 선정했다.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **데이터 1과 선정 이유** | **데이터 2와 선정 이유** | **데이터 3과 선정 이유** |
| ***김희령*** | 서울시 공유 자전거 | 넷플릭스 |  |
|  |  |  |
| ***문규림*** | 통신사 고객 이탈 |  |  |
|  |  |  |
| ***이원재*** | 가장 많이 구독한 유투브 채널 | 한국 관광객 수 |  |
|  |  |  |
| ***이현지*** | 서울시 공유 자전거 | 항공사 승객 만족도 | 블랙 프라이데이 세일 정보 |
| 서울시 날씨 데이터와 함께 분석하면 좋을 것 같아서 | 승객 만족도를 조사해서 서비스 개선을 목표로 분석하고 싶어서 | 블랙 프라이데이 세일 기간 고객의 상품 구매와 구매 빈도를 분석해보고 싶어서 |

표 1 - 팀원이 선택한 데이터

상단 표의 데이터 중 가장 마음에 드는 데이터를 투표하여 데이터를 선택했다. 데이터 선택 과정에서 각자의 관심 분야, 데이터 분석 가능성과 난이도, 프로젝트 목표의 적합성 등을 고려했다. 선정된 데이터는 항공사 승객 만족도 데이터로, Kaggle의 “Airline Passenger Satisfaction” 데이터셋을 사용했다. 고객 정보가 담긴 컬럼과 만족도 유형 컬럼이 다양해서 변수 간 관계를 파악하고, 데이터의 패턴을 파악하기 쉽다고 판단했으며, 이를 통해 의미 있는 분석 결과를 도출할 수 있다고 생각한다.

#### 분석 데이터 요약

항공사 만족도 데이터는 다음과 같은 컬럼으로 구성된다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **분류** | **컬럼명** | **데이터 타입** | **데이터 설명** | **비고** |
| **고객 정보** | id | Int | 고객 고유 id |  |
| Gender | Str | 성별 | Male/Female |
| Customer Type | Str | 충성/비충성 고객 | Loyal/Disloyal |
| Age | Int | 연령 |  |
| Type of Travel | Str | 여행 목적 | Personal/Business |
| Class | Str | 탑승 클래스 | Business/Eco/Eco plus |
| Flight Distance | Int | 비행 거리 | 이상치 존재 가능 |
| **기내 서비스 만족도** | Inflight wifi service | Int | 기내 와이파이 만족도 | 0~5점 |
| Food and drink | Int | 식음료 만족도 | 0~5점 |
| Seat comfort | Int | 좌석 편안함 | 0~5점 |
| Inflight entertainment | Int | 기내 오락 만족도 | 0~5점 |
| Leg room service | Int | 레그룸 서비스 만족도 | 0~5점 |
| Cleanliness | Int | 기내 청결도 | 0~5점 |
| Inflight service | Int | 기내 서비스 만족도 | 0~5점 |
| **탑승 절차 서비스 만족도** | Online boarding | Int | 온라인 보딩 만족도 | 0~5점 |
| Checkin service | Int | 체크인 서비스 만족도 | 0~5점 |
| On-board service | Int | 온보드 서비스 만족도 | 0~5점 |
| Ease of Online booking | Int | 온라인 예약 편의성 | 0~5점 |
| **이동 및 도착 만족도** | Baggage handling | Int | 수하물 처리 만족도 | 0~5점 |
| Gate location | Int | 게이트 위치 만족도 | 0~5점 |
| Departure/Arrival time convenient | Int | 출발/도착시간 편의성 | 0~5점 |
| Departure Delay in Minutes | Int | 출발 지연 시간(분) | 이상치 존재 가능 |
| Arrival Delay in Minutes | Int | 도착 지연 시간(분) | 이상치 존재 가능 |
| **만족도 결과** | satisfaction | Str | 만족/불만족 | Satisfaction / neutral or dissatisfaction |

Table 2 - 원본 데이터 설명

고객 관련 데이터로는 고객id, 성별, 고객 종류, 연령, 여행 목적, 탑승 클래스, 비행 거리 컬럼이 있다. 고객id, 성별, 고객 종류, 여행 목적은 명목형 데이터, 연령, 비행 거리는 연속형 데이터에 해당한다. 탑승 클래스는 순서형 데이터에 속한다. 기내 서비스 만족도, 탑승 절차 만족도, 이동 및 도착 만족도 범주에 해당하는 만족도 항목들은 전부 0점부터 5점 사이의 정수 값을 가지고 있는 순서형 데이터이다. 종합 만족도는 만족/불만족의 두 범주의 값으로 나타난다.

#### 프로젝트 목적과 배경

항공사는 고객의 티켓 예약, 발권, 수하물 처리, 기내 서비스의 전반적인 서비스를 제공한다. 항공사 고객 만족도 데이터에는 위 내용과 관련된 만족도 항목이 세분화 되어있다. 고객 만족도는 항공사의 평판과 성과에 직접적인 영향을 미치므로, 항공사는 고객 만족도를 높이기 위해 다양한 노력을 기울여야 한다. 고객의 기대 수준이 높아짐에 따라서 기본적인 서비스를 제공하는 것으로는 충분하지 않기 때문에, 고객을 세분화하여 차별화된 서비스를 제공해야 한다.

이 보고서에서는 차별화된 서비스 제공의 필요성에 따라서 항공사 서비스 만족도가 낮은 항목을 분석하고, 고객 세그먼트 별 만족도 수준을 파악한다. 이를 바탕으로 더 효과적인 서비스 개선 방안을 도출한다. 먼저 2. EDA를 통해 전체적인 데이터를 살피고 분류 기준을 선정한다. 그 후 3. 고객 세그먼테이션을 진행하고 그룹별로 만족도 항목을 분석한다.

### 2. EDA 수행 과정

#### 2.1 데이터 준비

- 데이터셋 전처리

원본 데이터인 Kaggle의 “Airline Passenger Satisfaction” 데이터는 만족도를 기반으로 고객의 만족도 결과 컬럼에 해당하는 satisfaction을 예측하는 목적으로 구성된다. Train.csv, test.csv 2개의 데이터로 구성되어 있으며, 두 데이터셋을 병합하여 airline\_passenger\_satisfaction.csv 파일을 생성했다.

두 파일 모두 clean up 데이터로, 결측치는 존재하지 않는다. 또한, 인구통계학적 특성을 반영하여 연속형 데이터인 연령을 범주화하여 0~9세 그릅부터 60세 이상까지의 총 7그룹으로 나누었다. 이 때 0~9세 고객의 만족도 데이터는 전체 약 12만 개 중 2190개로 매우 적은 수를 나타내고, 연령대가 매우 어리므로 분석 대상에서 제외했다.

Mysql에서 sql쿼리 작성 시 컬럼에 띄어쓰기가 있을 경우, 컬럼명을 제대로 인식하지 못한다. 이는 대괄호([]) 나 억음 부호(``)를 씌워야 올바른 컬럼명을 가져올 수 있다. 따라서 컬럼 명 띄어쓰기로 인해 발생하는 불편함을 방지하기 위해서, 컬럼 명의 띄어쓰기는 ‘\_’ 언더바로 대체했다. 또한, ‘Gender’, ‘Customer’, ‘Type of travel’, ‘satisfaction’ 컬럼의 문자열 데이터는 대문자 한단어로 변경하거나 중복 단어를 제거하는 등 프로그래밍 시 편의성을 높이고자 했다. 전처리 후 데이터는 다음과 같다.

|  |  |  |
| --- | --- | --- |
|  | **변경 전** | **변경 후** |
| **각 컬럼명** | Inflight wifi service | Inglight\_wifi\_service |
| **Gender 컬럼 데이터** | Female/Male | F/M |
| **Customer 컬럼 데이터** | Loyal Customer/disloyal Customer | Loyal/disloyal |
| **Type of travel 컬럼 데이터** | Personal travel/Business travel | Personal/Business |
| **Satisfaction 컬럼 데이터** | Satisfaction/neutral or dissatisfaction | S/D |
| **나이 범주화** | 0~85 | 10~19, … 50~59, 60+ |
| **0~9세 고객 데이터** | 분석 대상에서 제외 | |

Table 3 - 전처리 후 데이터 설명

#### 2.2 항공사 고객 만족도 데이터 EDA

먼저 연령, 성별과 같은 인구통계학적 특성을 기반으로 그룹을 나누어 데이터를 살펴본 후 각 컬럼 별 특성을 확인한다. 그 후 각 컬럼 간의 상관계수를 분석한다.

##### -기술 통계량 분석

표 1 - 고객 연령과 비행 거리 기술 통계량

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Age** | 127539 | 40.0034 | 14.6425 | 10 | 28 | 40 | 51 | 85 |
| **Flight\_Distance** | 127539 | 1196.0279 | 1001.6126 | 31.0 | 414.0 | 846.0 | 1747.0 | 4983.0 |

표 2 - 만족도 항목 별 기술 통계량

|  |  |  |  |
| --- | --- | --- | --- |
|  | **count** | **mean** | **std** |
| **Inflight\_wifi\_service** | 127539 | 2.7325 | 1.3321 |
| **Departure\_Arrival\_time\_convenient** | 127539 | 3.0507 | 1.5276 |
| **Ease\_of\_Online\_booking** | 127539 | 2.7608 | 1.4035 |
| **Gate\_location** | 127539 | 2.9775 | 1.2807 |
| **Food\_and\_drink** | 127539 | 3.2093 | 1.3288 |
| **Online\_boarding** | 127539 | 3.2656 | 1.3486 |
| **Seat\_comfort** | 127539 | 3.4505 | 1.3162 |
| **Inflight\_entertainment** | 127539 | 3.3649 | 1.3324 |
| **Onboard\_service** | 127539 | 3.3862 | 1.2872 |
| **Legroom\_service** | 127539 | 3.3535 | 1.3158 |
| **Baggage\_handling** | 127539 | 3.6310 | 1.1819 |
| **Checkin\_service** | 127539 | 3.3077 | 1.2659 |
| **Inflight\_service** | 127539 | 3.6413 | 1.1781 |
| **Cleanliness** | 127539 | 3.2919 | 1.3119 |

표1에서 전체 이용객 수는 127,539 명이고 고객의 평균 나이는 40세이다. 평균 비행 거리는 1196마일이고, 가장 짧은 비행은 31마일, 가장 긴 비행 거리는 4983마일이다.

표2에서 0점에서 5점 사이의 각 만족도 항목 별 평균과 표준편차를 나타낸다. 기내 와이파이 서비스 평균이 가장 낮으며, 기내 서비스 만족도가 가장 높고 표준편차가 가장 낮게 나타난다. 따라서 만족도 항목 중 기내 서비스 만족도의 분포가 평균에 가장 밀집되어 있다고 볼 수 있다.

##### - 승객 유형 별 분포

표 3 - 나이, 성별, 고객 유형, 클래스에 따른 승객 분포

|  |  |
| --- | --- |
|  |  |
|  |  |

좌측 상단의 나이 별 승객 수 그래프를 보았을 때, 전체 승객 유형 중 40대 이용객이 가장 많으며, 20대, 30대, 50대 순으로 비슷한 이용객 수를 보인다. 우측 상단의 성별에 따른 승객 수 분포에서는 여성 고객이 남성 고객보다 약간 더 많은 승객 수를 나타낸다. 좌측 하단의 고객 유형 수에서, 고객 유형이 충성 고객인 수가 비충성 고객보다 훨씬 많은 비중을 차지하고 있다. 우측 하단의 클래스별 승객 수 분포에서는 비즈니스 클래스 탑승객이 가장 많으며, 근소한 차이로 이코노미 클래스 탑승객이 뒤를 이룬다. 가장 저렴한 클래스인 이코노미 플러스 탑승객 수가 가장 적다.

|  |  |
| --- | --- |
|  | 텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진  자동 생성된 설명 |
|  | |
|  | |
|  | |

좌측 상단의 그래프는 출발/도착 지연 시간 산점도를 나타낸다. 출발 지연시간이 증가할 수록, 도착 지연시간도 증가하는 추세이다. 약 12만 7천 명의 승객 중, 출발/도착 지연을 겪지 않은 고객은 5만 4738명이다. 2시간 이상 출발 지연을 겪은 승객은 3029명으로 나타났다. 우측 상단의 여행 목적 별 그래프에서 출장 목적의 승객의 수가 개인 여행 목적 승객 수보다 많은 것으로 보인다. 그러나 출장 목적 승객과 달리, 개인 여행 목적 승객의 경우 만족 비율이 매우 낮다는 것을 알 수 있다. 중앙의 그래프는 탑승 클래스별 비행 거리 box plot과 거리 구간 별 고객 수 나타낸다. 주로 장거리 여행시 다른 클래스보다 비즈니스 클래스 탑승을 선호하는 것을 알 수 있다. 이코노미와 이코노미 플러스 클래스의 boxplot을 보면 1000마일 내외의 좁은 범위의 box를 가지고 있다. 이는 두 클래스가 1000마일 이내의 비행에 몰려있음을 나타낸다. 마지막 그래프는 비행 거리 구간 별 평균 만족도로, 1300 마일을 기준으로 평균 만족도가 급상승하는 것을 볼 수 있다. 중앙의 그래프와 함께 해석하면, 1000마일 내외의 고객층인 이코노미와 이코노미 플러스 클래스 승객들은 대부분 종합 만족도 비율이 낮고, 두 클래스에 비해 좀 더 장거리 여행을 하는 비즈니스 클래스 승객들은 종합 만족도 비율이 높다고 할 수 있다.

##### - 만족도 유형별 점수 분포

|  |
| --- |
| 텍스트, 도표, 스크린샷, 사각형이(가) 표시된 사진  자동 생성된 설명 |

다음은 총 14가지의 항목의 점수 분포를 boxplot으로 나타낸 것이다. 각 boxplot 내에서 좌측 box는 종합 만족도(satisfaction)에서 만족하지 않음을 선택한 고객이고, 우측 box는 만족한 고객이다. Baggage handling과 Inflight service 항목의 경우, 두 고객층의 box가 좁고 높은 점수에 가깝기 때문에 다른 항목에 비해 높은 점수 분포를 보인다. 대체적으로 만족한 고객의 만족도 유형 별 점수 분포는 만족하지 않은 고객의 분포보다 높은 점수대의 분포를 가지고 있다. Departure/Arrival time convenient 항목의 경우, 만족/불만족 고객 그룹의 점수 분포가 비슷하며, 점수대가 고르게 나타남을 알 수 있다.

##### -상관 관계 분석

상관계수 연산을 위해, 문자열/범주형 데이터를 정수형 데이터로 변환하는 과정을 수행했다. 종합 만족도가 불만족이면 0, 만족이면 1로, age\_group은 0부터 연령대 순서로 번호를 붙였다. Class는 저렴한 이코노미 플러스를 0, 이코노미를 1, 비즈니스를 2로 변환했다. Customer\_Type에서 충성고객은 1, 비충성 고객은 0으로 변환했고, Type of Travel에서 출장 목적은 1, 개인 여행목적은 0으로, Gender에서 여성은 0, 남성은 1로 변환했다. 변환 코드는 다음과 같다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

1. 고객 정보 컬럼과 satisfaction의 상관관계

|  |
| --- |
|  |

다음은 고객 정보(성별, 고객 유형, 연령대, 여행 목적, 클래스, 여행 거리)와 종합 만족도(satisfaction)에 따른 상관계수 행렬을 나타낸다. 성별과 종합 만족도의 상관계수는 0.01로, 관계성이 거의 없음으로 나타났다. 여행 목적, 클래스와 종합 만족도의 상관계수는 0.45로 강력한 상관관계가 있는 것으로 나타났다. 비행 거리, 승객 유형, 연령대 순으로 종합 만족도와의 관계성이 드러난다.

또한 고객 정보 간의 상관계수를 살펴보면, Type\_of\_Travel 과 Customer\_Type은 음의 상관관계가 나타난다. 이는 충성 고객이 주로 개인 여행을 목적으로 항공사를 이용하며, 비충성 고객은 주로 출장 목적으로 항공사를 이용하는 경향이 있다고 볼 수 있다. 그러나 표3의 충성/비충성 고객의 비중을 보았을 때, 충성고객이 비충성고객에 비해 압도적인 승객 수를 차지하므로 위와 같은 가설은 적절하지 않다. 실제 데이터의 고객 수는 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **충성(1)** | **비충성(0)** | 소계 |
| **출장(1)** | 61,766 | 21,877 | 83,643 |
| **개인여행(0)** | 35,739 | 180 | 35,919 |
| 소계 | 97,505 | 22,057 | 119,562 |

개인 여행목적의 비충성 고객의 수가 출장여행 목적의 비충성 고객 수 보다 매우 적기 때문에 다음과 같은 음의 상관계수가 나왔다고 추정할 수 있다.

다음으로 age\_group과 Customer\_Type은 양의 상관관계가 나타난다. 이는 고객의 연령대가 증가할수록, 충성고객이 많아진다는 것을 의미한다. 실제 데이터의 고객 수는 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **충성(1)** | **비충성(0)** | 충성 고객 비율(%) |
| **10대(0)** | 7563 | 1666 | 81.94 |
| **20대(1)** | 13171 | 11319 | 53.78 |
| **30대(2)** | 18901 | 5206 | 78.40 |
| **40대(3)** | 25308 | 2506 | 90.99 |
| **50대(4)** | 21635 | 857 | 96.18 |
| **60대 이상(5)** | 10925 | 501 | 95.61 |

고객 수가 가장 적은 연령대인 10대를 제외하면, 20대부터 고객 연령대가 증가할수록 충성 고객 비율이 점점 증가하는 것을 알 수 있다.

Class와 Type\_of\_Travel의 상관계수는 0.48로, 위의 상관 행렬 중 가장 강한 양의 관계를 보인다. 이는 비즈니스 클래스일수록 출장 목적의 여행 경향을 보인다고 할 수 있다. 실제 데이터 수는 다음과 같다.

|  |  |  |
| --- | --- | --- |
|  | **출장 목적(1)** | **개인 여행 목적(0)** |
| **비즈니스(2)** | 55569 | 2401 |
| **이코노미(1)** | 23546 | 29483 |
| **이코노미 플러스(0)** | 4526 | 4033 |

출장 목적의 비즈니스 클래스 승객이 개인여행 목적의 비즈니스 클래스 승객에 비해 압도적인 비중을 차지하고 있다. 그에 비해 이코노미, 이코노미 플러스 클래스는 두 여행 목적 승객 수 차이가 비즈니스 클래스에 비해 작게 나타난다. 비즈니스 클래스 승객의 수가 매우 많기 때문에 다음과 같은 상관계수가 나왔다고 볼 수 있다.

Flight\_distance와 Class의 상관계수는 0.43으로 강한 양의 상관관계가 나타난다. 이는 장거리 비행일수록, 비즈니스 클래스를 선호한다고 볼 수 있다. 탑승 클래스 별 평균 비행 거리는 다음과 같다.

|  |  |  |
| --- | --- | --- |
| **비즈니스(2)** | **이코노미(1)** | **이코노미 플러스(0)** |
| 1677.51 | 741.48 | 742.78 |

이코노미 클래스와 이코노미 플러스 클래스의 평균 비행거리는 거의 비슷하게 나타난다. 이는 항공사 정책에 따라 이코노미와 이코노미 플러스 클래스의 항공편이 비슷할 것이라고 해석된다.

2. Satisfaction과 만족도 항목 상관관계 행렬

|  |
| --- |
| 텍스트, 스크린샷이(가) 표시된 사진  자동 생성된 설명 |

다음은 0~5점의 점수를 가진 14가지의 만족도 항목과 종합 만족도인 satisfaction의 상관계수 행렬을 나타낸다. 출/도착 지연시간 만족도와 게이트 위치 편의성을 제외한 나머지 항목들은 종합 만족도와 양의 선형 관계를 나타낸다. Online\_boarding 만족도의 상관계수는 0.5로, 종합 만족도와 가장 강한 양의 관계를 보여준다.

Ease\_of\_Online\_booking 항목에서 높은 점수를 준 고객들은 Inflight\_wifi\_service 항목에도 높은 점수를 준 것으로 나타났다. 또한, Cleanliness, Food\_and\_drink, Seat\_comfort, Inflight\_entertainment 항목들은 강한 양의 선형관계를 보여 네 항목 모두 높은 점수를 주는 경향이 나타난다. 그러나 위 항목만으로는 전체 고객의 종합 만족도 상관관계를 판단하기 어려우므로, 고객을 세분화하고 분석을 통해 의미 있는 통찰력을 발견하는 것이 적절하다.

### 3.고객 세그멘테이션 및 그룹별 세부 분석

#### 3.1 상위 10개 고객 그룹 선정

본 분석에서는 고객 세분화를 위해 인구통계학적 특성을 기준으로 설정하였다. 인구통계학적 특성은 고객의 행동 패턴과 만족도에 영향을 미치는 중요한 요인으로, 이를 통해 고객을 보다 체계적으로 분류하고 분석할 수 있다. 특히, 고객의 연령대, 고객 타입, 클래스, 여행 목적을 기준으로 분류하였다.

상관계수 분석을 통해 성별이 다른 변수들과 거의 상관관계가 없음을 확인하였고, 따라서 성별은 분류 기준에서 제외하였다. 이러한 기준을 통해 고객을 72개의 그룹으로 세분화하였다.

이 중 상위 10개 고객 그룹을 선정한 결과, 이들 그룹은 전체 이용객의 60%를 차지하는 것으로 나타났다. 이는 상위 10개 그룹이 주요 고객층임을 시사한다. 주요 고객층을 대상으로 보다 구체적인 분석을 실시함으로써, 서비스 개선과 마케팅 전략 수립에 유용한 인사이트를 도출할 수 있을 것이다.

고객의 연령대, 고객 타입, 클래스, 여행 목적을 기준으로 나눈 고객 세그멘테이션을 통해 상위 10개 고객 그룹을 선정하고, 14가지 만족도 항목에 대해 각 그룹별로 세부 분석을 수행하였다. 이를 통해 각 고객 그룹의 특성과 요구사항을 명확히 파악하고, 맞춤형 서비스 제공 및 고객 만족도 향상을 위한 전략을 수립하는 데 기여하고자 한다.

4.1 회고