User

이 데이터 셋 안의 데이터를 살펴봐줘. 1) 어떤 열(column)이 있고, 각 열에 해당하는 변수는 뭘 뜻하는지 2) 데이터의 사이즈는 어떤지, 데이터 타입은 무엇인지 3) 학습과 테스트 데이터는 어떻게 분리가 되어 있는지 4) 결측 값은 어느 정도 인지 5) 라벨이 있는 데이터라면 분포는 어떻게 되어있는지 6) 데이터의 특이점(이상치)이 있는지

ChatGPT

이 데이터 세트는 타이타닉호 승객 명단에 관한 데이터로 보입니다. 다음은 데이터 세트에 포함된 각 열(column)과 그 의미입니다:

1. **PassengerId**: 승객 고유 번호
2. **Survived**: 생존 여부 (0 = 사망, 1 = 생존)
3. **Pclass**: 티켓 클래스 (1 = 1등석, 2 = 2등석, 3 = 3등석)
4. **Name**: 승객 이름
5. **Sex**: 성별
6. **Age**: 나이
7. **SibSp**: 함께 탑승한 형제자매나 배우자의 수
8. **Parch**: 함께 탑승한 부모님 또는 자녀의 수
9. **Ticket**: 티켓 번호
10. **Fare**: 운임
11. **Cabin**: 객실 번호
12. **Embarked**: 승선한 항구 (C = Cherbourg, Q = Queenstown, S = Southampton)

이제 데이터의 사이즈, 타입, 학습과 테스트 데이터 분리 여부, 결측치, 라벨 분포, 특이점을 추가로 분석해 보겠습니다. ​​

**데이터 세트 분석 결과:**

**데이터 구조 및 타입**

* 총 891개의 엔트리가 있으며, 각 엔트리는 12개의 열로 구성되어 있습니다.
* 데이터 타입은 정수형(int64), 실수형(float64), 그리고 객체(object) 타입이 포함되어 있습니다.

**결측치 분석**

* **Age**: 177개의 결측치가 있습니다.
* **Cabin**: 687개의 결측치가 있습니다, 즉 대부분의 데이터에서 객실 번호가 누락되어 있습니다.
* **Embarked**: 2개의 결측치가 있습니다.

**라벨 분포**

* 생존 여부 (Survived)는 0과 1의 값을 가집니다. 여기서 0은 사망을, 1은 생존을 의미합니다.
  + 사망한 승객: 549명
  + 생존한 승객: 342명

**데이터 특이점 (이상치)**

* **Fare** (운임): 최대값이 512.3292로 매우 높은 비용을 나타내며, 평균보다 훨씬 높습니다. 이는 몇몇 이상치가 존재할 수 있음을 시사합니다.
* **Age**와 **SibSp** (동승 형제자매/배우자 수)에서도 표준편차가 상대적으로 큰 것을 볼 수 있습니다.

**학습 및 테스트 데이터 분리**

* 제공된 데이터에서 학습과 테스트 데이터의 분리 정보는 포함되어 있지 않습니다. 데이터 세트를 적절히 분리하여 모델 학습 및 평가에 사용할 필요가 있습니다. 데이터를 분할할 때 일반적으로 70-30, 80-20 또는 교차 검증 방식을 사용할 수 있습니다.

이 데이터를 사용하여 분석이나 머신 러닝 모델 개발을 계속 진행하려면 추가적인 데이터 정제 작업이 필요할 것입니다. 데이터의 어떤 부분을 더 자세히 알아보고 싶으신가요? ​​

User

알고 싶은 결과(Survived 여부)를 Y로 설정하여 EDA 분석을 시작해줘. 1) 데이터 결측값 확인 2) Survived 에 대한 비율 확인 3) 각 열(column)을 X로 설정하고 각 X에 대한 변수 분포 및 비율 확인 4) 각 열(column)과 Survived 과의 상관관계 확인 EDA 분석한 결과를 보여줄 때 각 데이터 유형에 따라, 시각화 기법, 그래프 종류 중 가장 적절한 것을 잘 선택해서 하나씩만 보여줘. - 데이터 유형 : 수치형(연속형/이산형), 범주형(순서형, 명목형) - 시각화 기법 : 일변량 비시각화 기법/일변량 시각화 기법/ 다변량 비시각화 기법 / 다변량 시각화 기법 - 사용 라이브러리 : pandas/seaborn - 그래프 종류 : 히스토그램/커널밀도추정함수/분포도/러그플롯/막대플롯/포인트플롯/박스플롯/바이올린 플롯/카운트 플롯/파이/히트맵/라인플롯/산점도/회귀선포함산점도 등

ChatGPT

**EDA 분석 결과**

**1. 데이터 결측치 확인**

* **Age**: 177개의 결측치
* **Cabin**: 687개의 결측치
* **Embarked**: 2개의 결측치

**2. Survived의 분포**

* 생존자: 38.4%
* 사망자: 61.6%

**3. 각 열에 대한 변수 분포 및 비율**

* **수치형 데이터 (Age, Fare, SibSp, Parch)**:
  + **Age**: 평균 약 29.7세, 범위 0.42세(유아)에서 80세
  + **Fare**: 평균 약 32.2, 범위 0에서 512.33
  + **SibSp** 및 **Parch**는 대부분의 승객이 혼자 여행하거나 가족 없이 여행했음을 나타냅니다.
* **범주형 데이터 (Name, Sex, Ticket, Cabin, Embarked)**:
  + **Sex**: 남성 약 64.8%, 여성 약 35.2%
  + **Embarked**: S항에서 승선한 승객이 약 72.3%, C항 약 18.9%, Q항 약 8.6%

**4. 각 열과 Survived의 상관관계**

* **Pclass**와 **Fare**는 Survived와 가장 강한 상관관계를 보입니다. 특히, Pclass가 낮을수록 (1등석), Fare가 높을수록 생존률이 높았습니다.
* **Age**는 상대적으로 낮은 음의 상관관계를 보여, 나이가 많을수록 생존 확률이 약간 낮았음을 나타냅니다.

**시각화**

* 결측치는 막대 그래프로 시각화하여 각 변수의 누락 데이터 수를 보여주었습니다.
* Survived의 분포는 파이 차트로 나타내 생존과 사망 비율을 시각적으로 표현하였습니다.
* Survived와 다른 변수들과의 상관관계는 히트맵으로 표현하여 각 변수 간의 상관계수를 색상으로 구분하였습니다.

추가적으로 필요한 분석이나 궁금한 사항이 있으시면 알려주세요. ​​

User

EDA를 통해 나온 사실을 정리하고 가설을 세워줘. 그리고 너가 역사학자의 시선에서 해당 분야의 전문가라고 생각하고 더욱 날카로운 가설을 생각해줘. 처음의 가설과 새로운 가설을 구분해서 정리해줘.

ChatGPT

**기존의 EDA 분석을 통해 도출된 사실**

1. **결측치가 많은 데이터**: Age와 Cabin 열에 상당한 결측치가 존재합니다. 이는 데이터 분석시 정확성에 영향을 줄 수 있습니다.
2. **생존율**: 전체 승객 중 약 38.4%가 생존했습니다. 이는 타이타닉의 참사에서 상대적으로 적은 수의 생존자가 있었다는 역사적 사실과 일치합니다.
3. **성별과 생존율**: 남성이 전체 승객의 약 64.8%를 차지하며, 여성은 약 35.2%입니다. 일반적으로 여성과 어린이의 생존율이 더 높았다는 보고가 있습니다.
4. **클래스와 생존율**: 더 높은 클래스(1등석)의 승객이 더 높은 생존율을 보였습니다. 이는 더 높은 요금을 지불한 승객이 더 나은 구조 기회를 가졌을 수 있음을 시사합니다.
5. **상관관계 분석**: Pclass와 Fare는 Survived와 강한 상관관계를 보였으며, 특히 Pclass가 높을수록 생존 확률이 높았습니다.

**초기 가설 설정**

* **가설 1**: 나이가 많은 승객은 나이가 적은 승객에 비해 생존률이 낮을 것이다.
* **가설 2**: 1등석 승객은 2등석이나 3등석 승객보다 생존 확률이 높을 것이다.
* **가설 3**: 객실 위치에 따라 생존율이 달라질 것이다. 특히, 더 쉽게 탈출할 수 있는 위치에 있는 객실의 승객들이 더 많이 생존했을 것이다.

**역사학자의 시각에서의 추가 가설**

* **가설 4 (역사적 배경)**: 타이타닉의 구조 작업 중 "여성과 어린이 먼저"라는 원칙이 적용되었기 때문에, 여성과 어린이의 생존율이 남성보다 훨씬 높을 것이다.
* **가설 5 (사회적 계층)**: 1등석 승객들이 사회적, 경제적 지위가 높았기 때문에, 구조 작업에서 우선적으로 고려되었을 것이다. 이는 1등석 승객의 높은 생존율로 이어졌을 것이다.
* **가설 6 (승선 위치의 영향)**: 승선 위치(Embarked)에 따라 승객들의 구성이 다르며, 이는 생존율에 영향을 미쳤을 수 있다. 예를 들어, 일부 항구에서 탑승한 승객들이 더 높은 사회경제적 지위를 가지고 있을 가능성이 있으며, 이는 그들의 생존율에 긍정적인 영향을 미쳤을 것이다.

이러한 가설은 데이터 분석을 통해 검증할 수 있으며, 더욱 정교한 분석을 통해 이 가설들의 타당성을 평가할 수 있습니다. 추가적인 데이터 분석이나 특정 가설에 대한 검증이 필요하시면 말씀해 주세요.

User

앞선 대화를 모두 정리해서 ‘Survived에 영향을 미치는 요소 EDA 분석 및 시각화’ 보고서를 작성해줘. 특히, 시각화 결과물도 반드시 보고서에 포함해줘.

ChatGPT

**Survived에 영향을 미치는 요소 EDA 분석 및 시각화 보고서**

**1. 데이터 세트 개요**

이 데이터 세트는 타이타닉호의 승객 명단에 관한 정보를 포함하고 있으며, 승객의 생존 여부를 중심으로 다양한 특성을 포함하고 있습니다. 데이터는 총 891명의 승객에 대한 정보와 다음의 열을 포함하고 있습니다:

* **PassengerId**: 승객의 고유 식별 번호
* **Survived**: 생존 여부 (0 = 사망, 1 = 생존)
* **Pclass**: 승객의 티켓 클래스
* **Name**: 승객의 이름
* **Sex**: 승객의 성별
* **Age**: 승객의 나이
* **SibSp**: 함께 탑승한 형제자매나 배우자의 수
* **Parch**: 함께 탑승한 부모님 또는 자녀의 수
* **Ticket**: 티켓 번호
* **Fare**: 티켓의 요금
* **Cabin**: 객실 번호
* **Embarked**: 탑승한 항구

**2. 데이터 정제 및 결측치 처리**

데이터 세트에는 일부 열에서 결측치가 발견되었습니다. 주요 결측치는 다음과 같습니다:

* **Age**: 177개의 결측치
* **Cabin**: 687개의 결측치
* **Embarked**: 2개의 결측치

결측치의 처리는 분석의 정확도를 위해 중요합니다. 특히 Cabin 열의 결측치는 승객의 생존에 중요한 변수일 수 있으므로, 주의 깊게 다루어야 합니다.

**3. 분석 결과 요약**

**3.1 결측치 시각화**

결측치 분포를 나타내는 막대 그래프는 다음과 같습니다:

**3.2 Survived의 분포**

생존자와 비생존자의 비율을 나타내는 파이 차트는 다음과 같습니다:

**3.3 상관관계 분석**

Survived와 다른 변수들과의 상관관계를 나타내는 히트맵은 다음과 같습니다:

**4. 초기 가설 설정**

* **가설 1**: 나이가 많은 승객은 나이가 적은 승객에 비해 생존률이 낮을 것이다.
* **가설 2**: 1등석 승객은 2등석이나 3등석 승객보다 생존 확률이 높을 것이다.
* **가설 3**: 객실 위치에 따라 생존율이 달라질 것이다.

**5. 역사학자의 시각에서의 추가 가설**

* **가설 4**: "여성과 어린이 먼저"라는 원칙에 따라 여성과 어린이의 생존율이 높을 것이다.
* **가설 5**: 사회적, 경제적 지위가 높은 1등석 승객은 구조 작업에서 우선적으로 고려되어 높은 생존율을 보일 것이다.
* **가설 6**: 승선 위치에 따라 생존율에 차이가 있을 것이다.

**6. 결론 및 추천**

이 분석을 통해 타이타닉의 생존 데이터에 대한 깊은 이해를 얻을 수 있었습니다. 데이터의 정제와 가설 설정을 통해 추가 분석을 진행할 수 있는 기반을 마련하였습니다. 향후 분석에서는 더욱 세부적인 변수의 영향력을 분석하고, 결측치 처리 방법을 개선하여 데이터의 정확성을 높일 필요가 있습니다.

보고서의 시각 자료들은 분석 과정에서 얻은 중요한 인사이트를 효과적으로 전달하는 데 도움을 주었습니다. 추가적인 데이터 분석이나 가설 검증을 위한 진행이 필요할 경우, 적절한 방법론과 도구를 사용하여 보다 정확한 결과를 도출할 수 있을 것입니다.