

타이타닉 생존 분석 프로젝트 - 최종 보고서

1. 프로젝트 개요 및 과정

1.1 프로젝트 개요

타이타닉 생존 분석 프로젝트는 타이타닉 호의 승객 데이터셋을 이용하여 생존 여부를 예측하고, 생존에 영향을 미치는 요인을 분석하는 것을 목표로 합니다. 이 프로젝트는 데이터 전처리, 탐색적 데이터 분석(EDA), 그리고 머신러닝 모델링의 세 단계로 진행되었습니다.

1.2 프로젝트 진행 과정

1. 데이터 전처리 (01_data_preprocessing.ipynb)

- **결측치 처리:** 승객 연령, 객실 번호 등 여러 변수에 결측치가 존재하였으며, 이를 적절한 방식으로 대체하여 분석의 신뢰성을 높였습니다. 예를 들어, 연령의 결측치는 승객의 사회적 계층 및 성별을 고려하여 평균 또는 중앙값으로 채웠습니다.

- **코드 예시:**

```
# 결측치 처리 코드
df['Age'].fillna(df.groupby(['Pclass', 'Sex'])
                 ['Age'].transform('median'), inplace=True)
```

- **이유:** 결측치를 적절히 처리함으로써 데이터의 신뢰성을 높이고 분석 과정에서 왜곡을 줄이기 위함입니다.
- **파생변수 생성:** 생존 여부에 영향을 줄 수 있는 추가적인 파생변수를 생성했습니다. 예를 들어, 가족 구성 여부를 나타내는 변수와 객실 등급을 분류한 변수를 추가하여 분석의 정밀도를 높였습니다.

- **코드 예시:**

```
# 가족 크기 파생변수 생성
df['FamilySize'] = df['SibSp'] + df['Parch'] + 1
```

- **이유:** 가족 크기는 생존율에 영향을 미치는 중요한 요인으로, 이를 변수로 추가하여 모델의 예측 성능을 높이기 위함입니다.
- **데이터 타입 최적화:** 메모리 사용을 줄이고 데이터 처리를 효율적으로 하기 위해, 변수의 데이터 타입을 최적화했습니다.

- **코드 예시:**

```
# 데이터 타입 최적화
df['Pclass'] = df['Pclass'].astype('int8')
```

- **이유:** 메모리 사용량을 줄이고 데이터 처리 속도를 개선하기 위함입니다.
- **전처리된 데이터 저장:** 전처리 후 데이터는 `processed/titanic_processed.csv` 파일로 저장되었습니다.
 - **코드 예시:**

```
# 전처리된 데이터 저장
df.to_csv('processed/titanic_processed.csv', index=False)
```

2. 탐색적 데이터 분석 (02_eda.ipynb)

- **기본 통계 분석:** 데이터셋의 주요 통계 정보를 분석하여 각 변수의 분포와 데이터 특성을 파악했습니다.
 - **코드 및 결과 예시:**
- **결과:** 주요 변수의 평균, 중앙값, 분산 등을 확인하여 데이터의 전반적인 특성을 파악했습니다.
- **변수 간 관계 분석:** 생존율에 영향을 미치는 주요 변수(예: 성별, 나이, 객실 등급 등) 간의 관계를 분석했습니다. 예를 들어, 성별에 따라 생존율이 크게 차이나는 것을 발견했습니다.
 - **시각화 예시:**

```
import seaborn as sns
import matplotlib.pyplot as plt

# 성별에 따른 생존율 시각화
sns.barplot(x='Sex', y='Survived', data=df)
plt.title('성별에 따른 생존율')
plt.show()
```

- **이유:** 시각화를 통해 변수 간의 관계를 명확하게 이해하고, 인사이트를 도출하기 위함입니다.
- **생존율 패턴 분석:** 다양한 그룹 간의 생존율을 비교 분석하여, 상위 객실 등급과 여성 승객이 상대적으로 높은 생존율을 보임을 확인했습니다.
 - **시각화 결과:** 상위 객실 등급 승객이 더 높은 생존율을 보였습니다. 이는 당시의 사회적 지위가 생존에 큰 영향을 미쳤음을 보여줍니다.

3. 머신러닝 분석 (03_ml_analysis.ipynb)

- **기본 모델 구현:** 로지스틱 회귀, 의사결정나무, 랜덤 포레스트 모델을 구현하여 생존 여부를 예측했습니다.

- **코드 예시:**

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

# 데이터 분할
X = df.drop('Survived', axis=1)
y = df['Survived']
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.2, random_state=42)

# 로지스틱 회귀 모델 학습
model = LogisticRegression()
model.fit(X_train, y_train)
```

- **이유:** 여러 모델을 비교하여 가장 성능이 좋은 모델을 선택하기 위해 기본적인 머신러닝 모델을 구현했습니다.
- **하이퍼파라미터 최적화:** 각 모델의 성능을 극대화하기 위해 하이퍼파라미터를 최적화했습니다. 특히 랜덤 포레스트의 경우, 그리드 서치를 이용해 최적의 파라미터를 찾았습니다.

- **코드 예시:**

```
from sklearn.model_selection import GridSearchCV

# 랜덤 포레스트 하이퍼파라미터 튜닝
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [4, 6, 8],
    'min_samples_split': [2, 5, 10]
}
grid_search = GridSearchCV(RandomForestClassifier(),
                           param_grid, cv=5)
grid_search.fit(X_train, y_train)
```

- **이유:** 모델 성능을 최대한으로 끌어올리기 위해 하이퍼파라미터를 최적화했습니다.
- **모델 성능 평가:** 정확도, 정밀도, 재현율 등의 지표를 사용해 모델을 평가하였으며, 랜덤 포레스트가 가장 우수한 성능(83.84% 정확도)을 보였습니다.

- **코드 예시:**

```
from sklearn.metrics import accuracy_score, precision_score,
recall_score
```

```
# 모델 예측 및 평가
y_pred = grid_search.best_estimator_.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f'랜덤 포레스트 정확도: {accuracy:.2f}')
```

- **결과:** 랜덤 포레스트 모델이 83.84%의 정확도를 보여 가장 우수한 성능을 기록했습니다.
- **특성 중요도 분석:** 모델이 예측하는 데 중요한 특성을 분석하여, 성별, 객실 등급, 요금, 나이 등이 생존 예측에 중요한 역할을 하는 것을 확인했습니다.
- **시각화 예시:**

```
# 특성 중요도 시각화
importances =
grid_search.best_estimator_.feature_importances_
features = X.columns
plt.figure(figsize=(10, 6))
sns.barplot(x=importances, y=features)
plt.title('랜덤 포레스트 특성 중요도')
plt.show()
```

- **이유:** 모델의 해석 가능성을 높이고, 중요한 특성을 파악하여 인사이트를 도출하기 위함입니다.

2. 주요 발견 사항

2.1 데이터 분석 결과

1. 생존율 패턴

- **전체 생존율:** 약 38.4%로, 전체 승객 중 약 3분의 1이 생존했습니다.
- **성별 차이:** 여성 생존율이 남성보다 현저히 높았습니다. 이는 당시 구조 작업에서 "여성과 아이 먼저"라는 원칙이 적용된 결과로 해석됩니다.
- **객실 등급:** 상위 객실 등급(1등급) 승객일수록 생존율이 높았습니다. 이는 사회경제적 지위가 생존 가능성에 큰 영향을 미쳤음을 보여줍니다.
- **가족 크기:** 가족 크기가 중간 규모(2-4명)일 때 생존율이 상대적으로 높았습니다. 이는 너무 큰 가족은 모두 구조되기 어렵고, 단독 승객은 도움을 받기 어려웠기 때문일 수 있습니다.

2. 사회경제적 요인

- **객실 등급과 요금:** 객실 등급과 승선 요금이 생존과 강한 상관관계를 보였습니다. 상위 계층의 승객일수록 생존 확률이 높았으며, 이는 당시 사회적 불평등이 생존 가능성에 반영된 결과로 볼 수 있습니다.
- **승선 항구:** 승선 항구별로 생존율에 차이가 있었습니다. 이는 특정 항구에서 승선한 승객들의 사회경제적 배경이 다르기 때문일 가능성이 있습니다.

2.2 머신러닝 분석 결과

1. 모델 성능 비교

- 로지스틱 회귀: 80.06% 정확도
- 의사결정나무: 82.02% 정확도
- 랜덤 포레스트: 83.84% 정확도 (최고 성능)

2. 주요 생존 결정 요인

- **성별:** 생존 예측에 가장 중요한 변수로 분석되었습니다.
- **객실 등급:** 높은 객실 등급일수록 생존 가능성이 높았습니다.
- **요금:** 승선 요금 역시 중요한 변수로 작용했습니다.
- **나이 및 가족 구성:** 나이와 가족 구성 여부도 생존 예측에 기여하는 변수였습니다.

3. 결론 및 시사점

3.1 최종 결론

1. 생존 예측 모델의 신뢰성

- 본 프로젝트에서 개발한 생존 예측 모델은 약 80% 이상의 정확도를 보여, 타이타닉 생존 여부를 비교적 신뢰성 있게 예측할 수 있음을 확인했습니다.

2. 핵심 인사이트

- "여성과 아이 먼저" 원칙이 실제로 적용되어 생존율에 큰 영향을 미쳤습니다.
- 사회경제적 지위가 생존 가능성에 결정적인 영향을 미쳤음을 데이터로 확인할 수 있었습니다.
- 가족이 함께 있는 것이 생존 확률을 높일 수 있는 중요한 요인으로 작용했습니다.

3.2 프로젝트 성과

1. 기술적 성과

- 체계적인 데이터 전처리 및 다양한 시각화 기법을 활용하여 데이터를 분석하였습니다.
- 머신러닝 모델을 활용해 생존 여부를 예측하고, 모델의 성능을 최적화하였습니다.

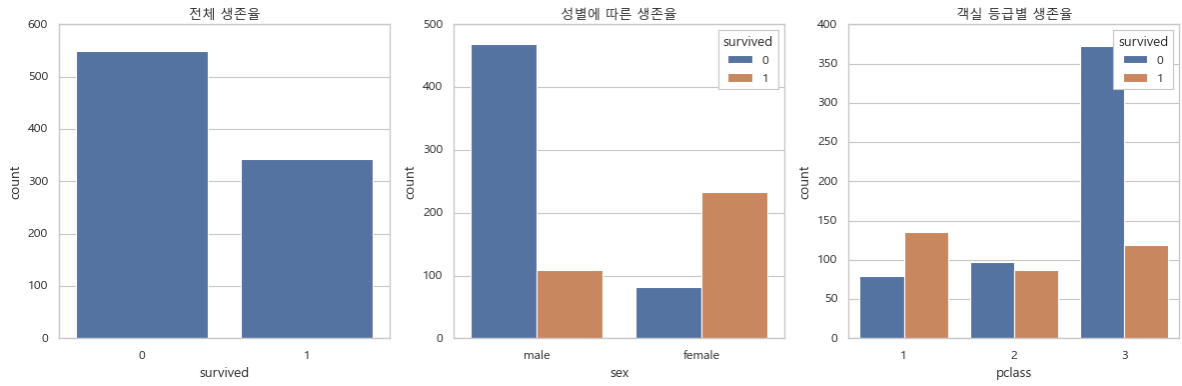
2. 분석적 성과

- 타이타닉 승객의 생존에 영향을 미치는 주요 요인을 식별하였으며, 이를 통해 타이타닉 사건에 대한 보다 깊이 있는 이해를 도출할 수 있었습니다.
- 데이터 기반의 분석 결과를 통해 역사적 사건에 대한 중요한 통찰을 제공하였습니다.

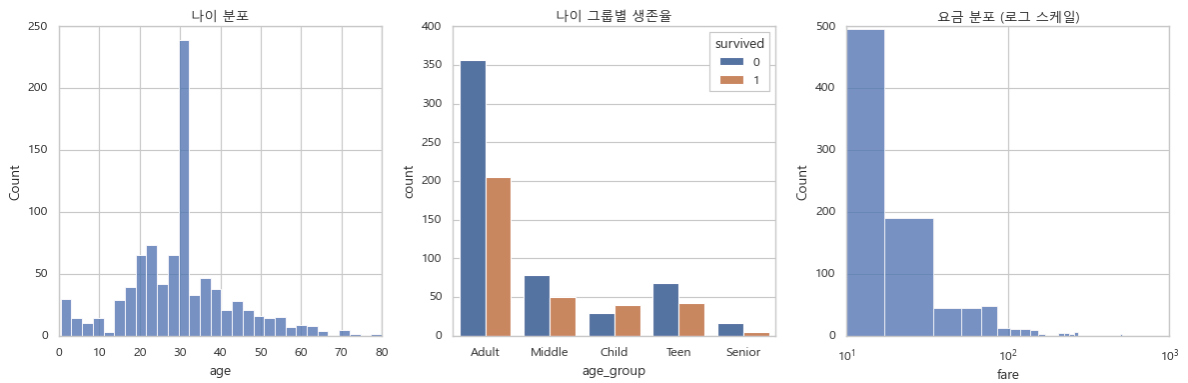
(부첨)데이터 시각화

기초 분석

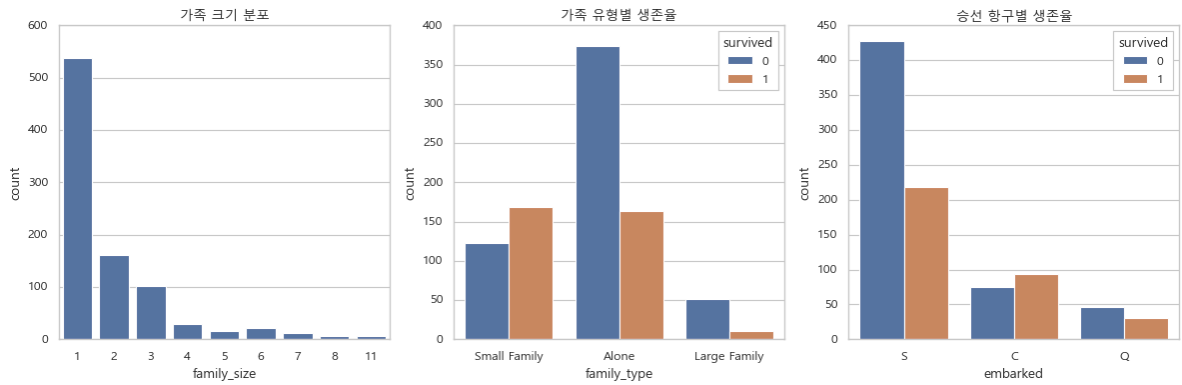
1. 생존율 기본 분석



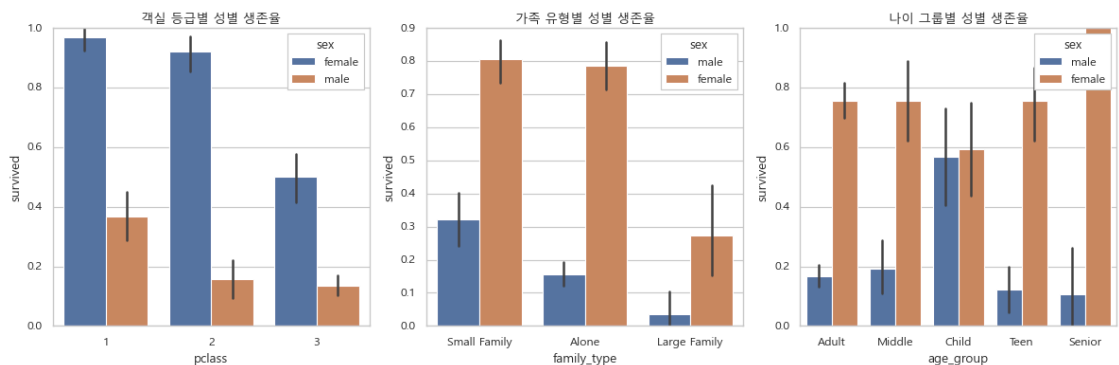
2. 나이와 요금 분포



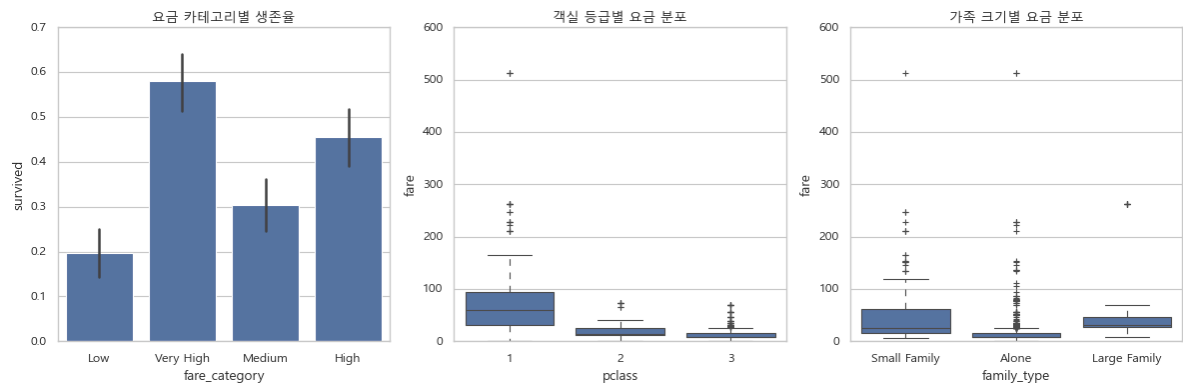
3. 가족 관련 분석



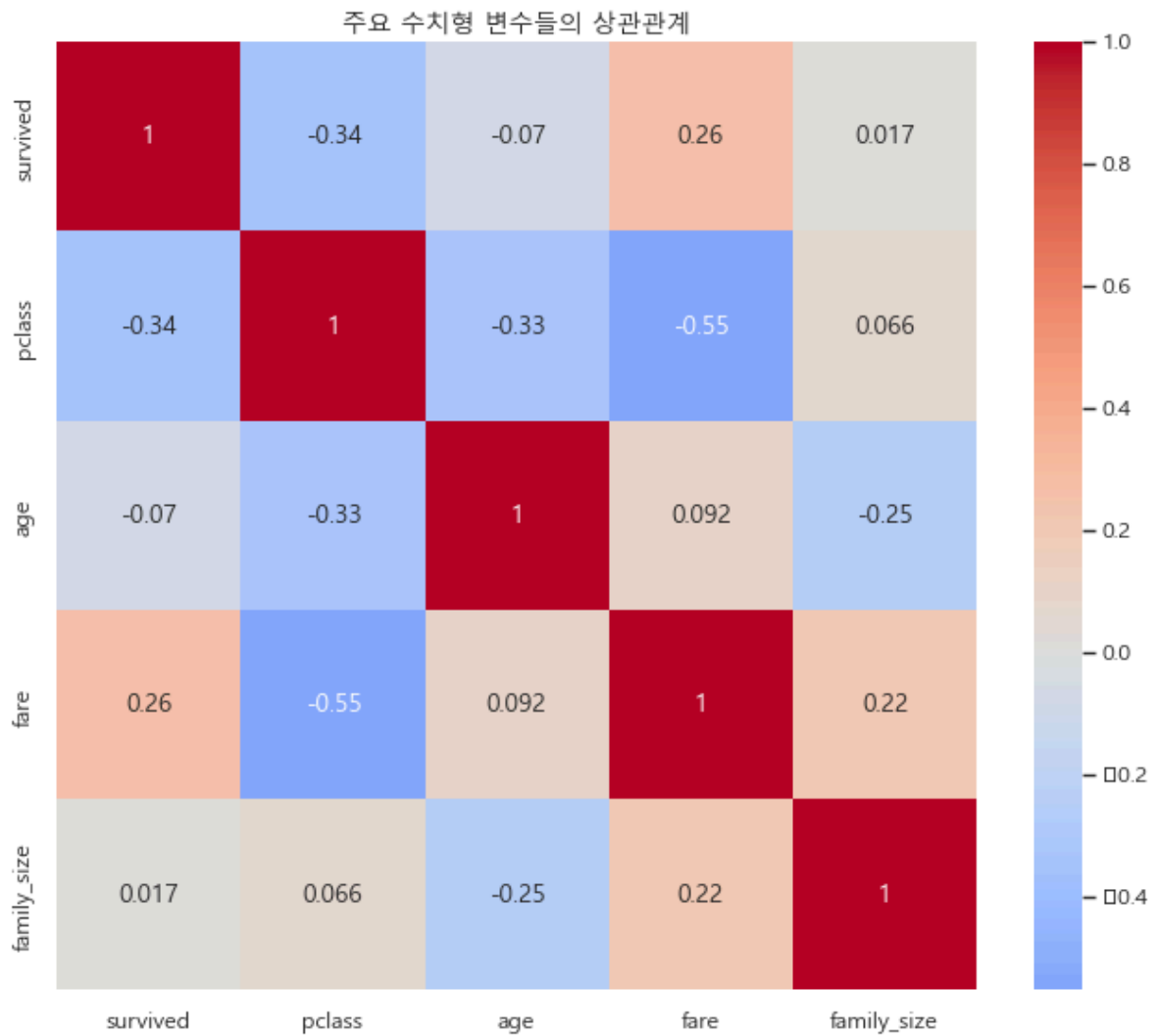
1. 성별, 객실 등급, 생존율의 관계



4. 요금과 생존율의 관계

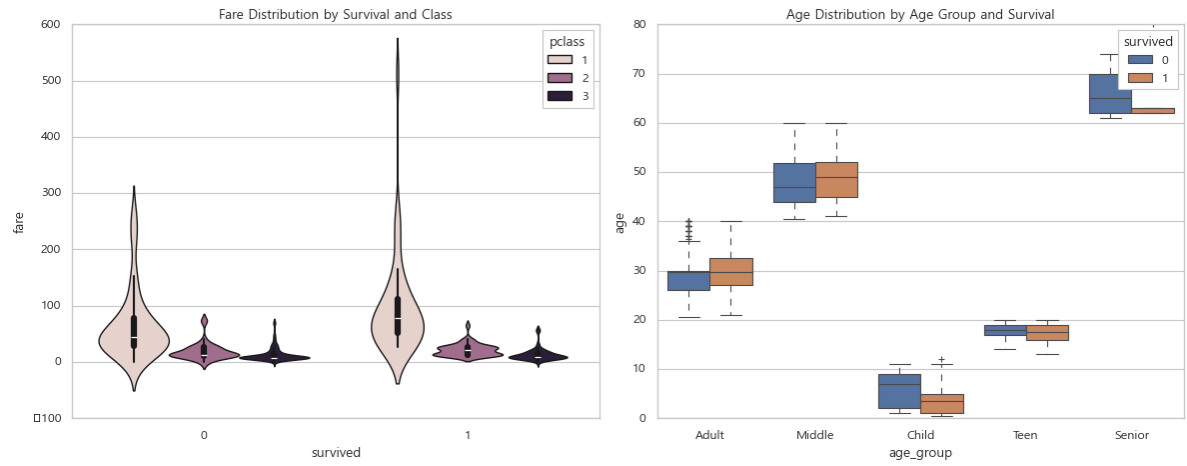


5. 주요 변수들의 상관관계

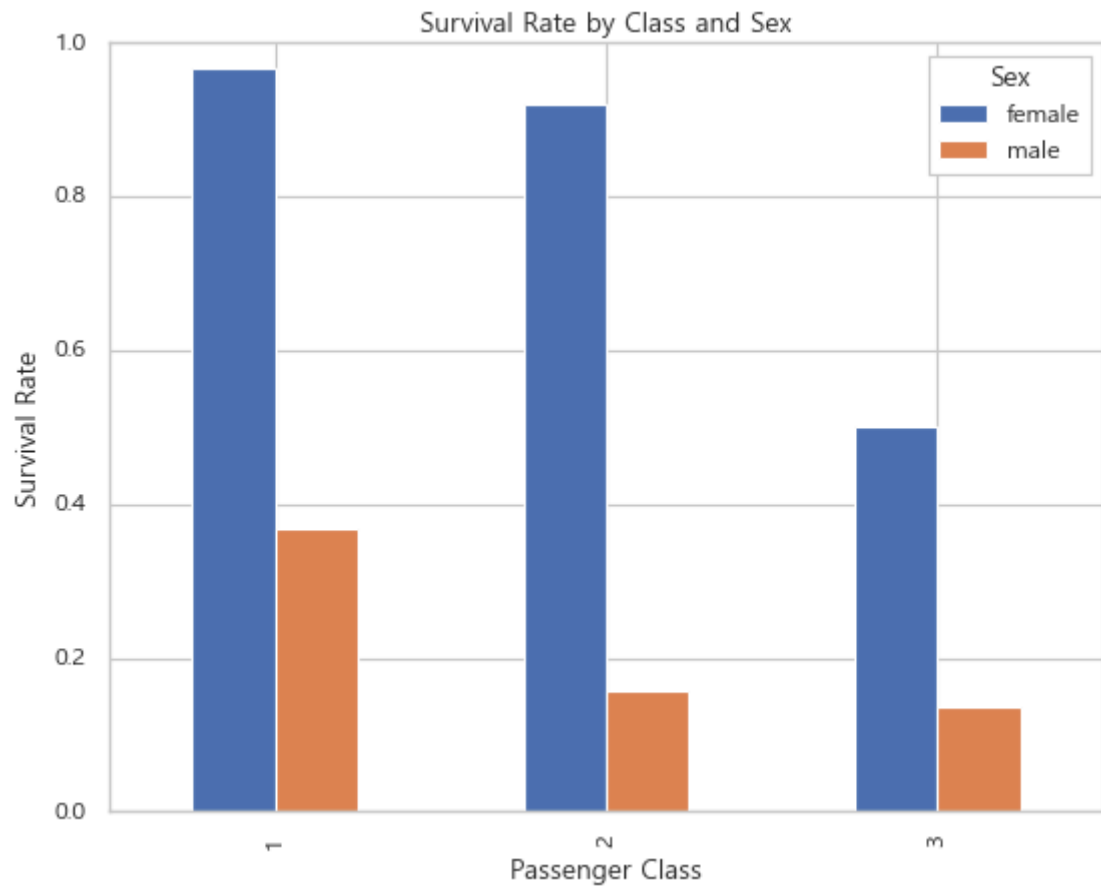


고급 시각화 분석

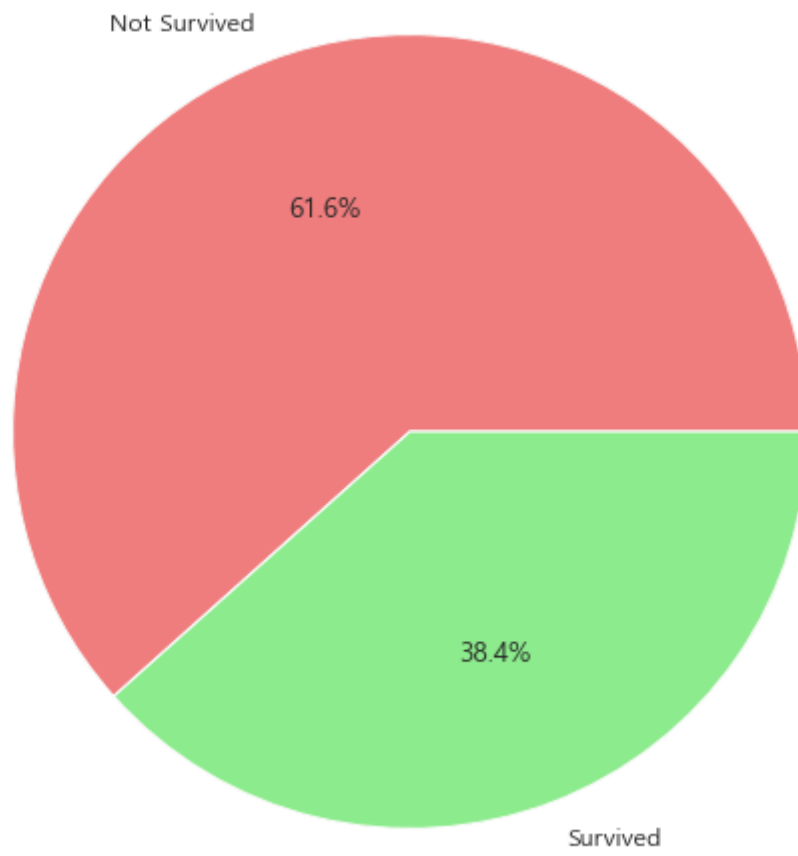
1. 바이올린 플롯과 박스플롯 조합



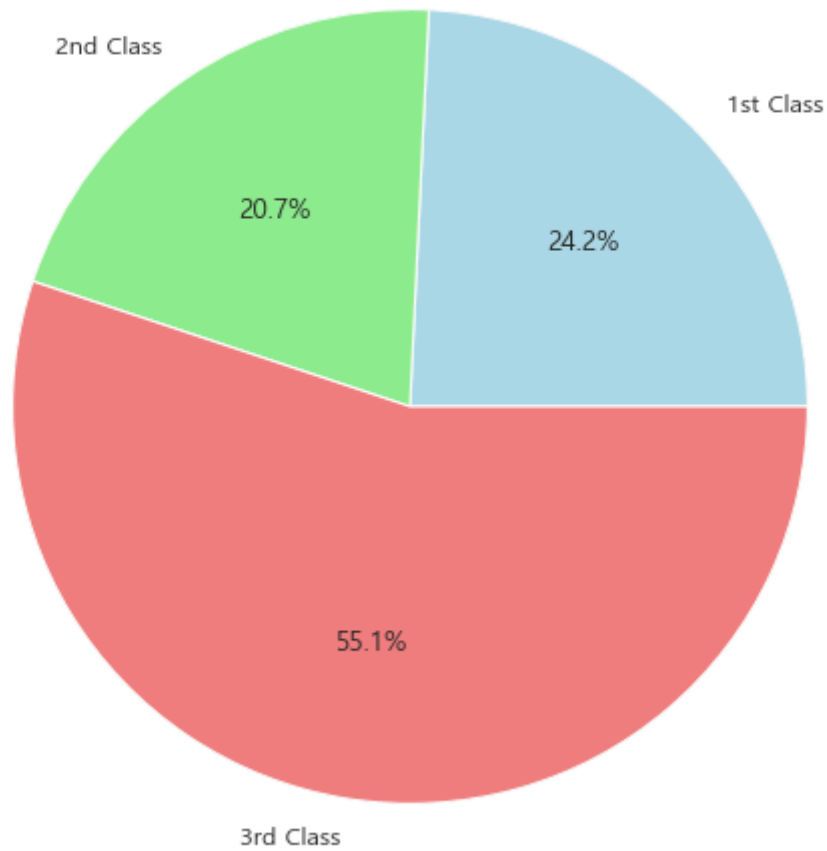
2. 복합 막대 그래프와 파이 차트



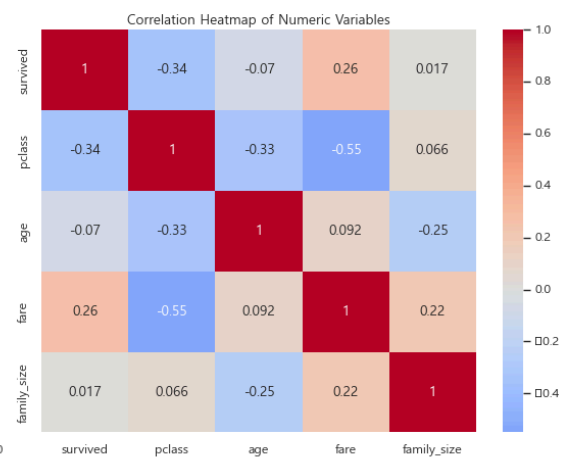
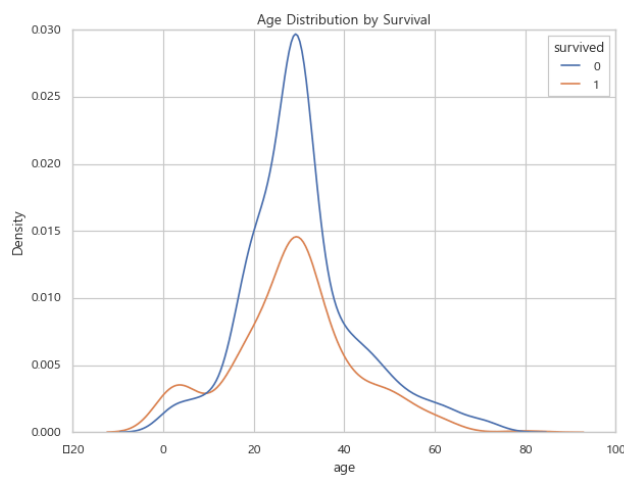
Overall Survival Distribution



Passenger Distribution by Class



2. KDE 플롯과 히트맵



GitHub : <https://github.com/kjhgreat/titanic-analysis/>