Check-In

이번 주 스터디 첫 모임을 위해 공부하며 느낀 점을 공유해주세요

1주차 데이터 전처리&EDA

전처리&EDA

- 정의 및 과정
- 문제 정의
- 데이터 탐색
- 결측치 처리
- 상관관계 파악

캐글 (Kaggle)

1주차 데이터 전처리&EDA

전처리&EDA

- 정의 및 과정
- 문제 정의
- 데이터 탐색
- 결측치 처리
- 상관관계 파악

캐글 (Kaggle)

EDA란?

탐색적 데이터 분석(Exploratory Data Analysis)의 줄임 말로 수집한
데이터를 분석하기 전에 **그래프나 통계적인 방법으로 자료를 직관적**으로 바라보는 과정

전처리&EDA

- 정의 및 과정
- 문제 정의
- 데이터 탐색
- 결측치 처리
- 상관관계 파악

캐글 (Kaggle)

1. 문제 정의

- 분석의 목적 확인
- 해결하고자 하는 문제 정의

=> 데이터 탐색 전 중요한 단계

ex) Titanic 생존률

2. 데이터 탐색

- 데이터 내 변수 확인
- 개별 변수의 이름이나 설명
- 이상치 및 결측치 확인

3. 이상치 처리

- 통계값 활용(Mean, Median)
- 시각화 활용
- 머신러닝 기법 활용

ex) 결측치 처리 5가지 기법

4. 상관관계 파악

- Categorical vs Numerical
- Categorical 해석 및 처리

ex) Embarked vs Age

전처리&EDA

정의 및 과정

문제 정의

데이터 탐색

결측치 처리

상관관계 파악

캐글 (Kaggle)

1. 문제 정의

- 분석의 목적 확인
- 해결하고자 하는 문제 정의

=> 데이터 탐색 전 중요한 단계

ex) Titanic 생존률

2. 데이터 탐색

- 데이터 내 변수 확인
- 개별 변수의 이름이나 설명
- 이상치 및 결측치 확인

EDA가 끝나면 이제 Modeling을 하게 된다!

3. 이상치 처리

- 통계값 활용(Mean, Median)
- 시각화 활용
- 머신러닝 기법 활용

ex) 결측치 처리 5가지 기법

4. 상관관계 파악

- Categorical vs Numerical
- Categorical 해석 및 처리

ex) Embarked vs Age

문제 정의

캐글 Competition의 목적 그대로 타이타닉 사고에서 어떤 승객이 살아남는지 예측하기

Titanic - Machine Learning from Disaster

데이터 탐색

1. head()와 tail()로 데이터의 전체 구조를 파악

In [3]:
```
training.head()
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

In [4]:
```
training.tail()
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q

데이터 탐색

2. info()는 데이터에 대한 설명

```
In [16]: testing.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 9 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  418 non-null    int64  
 1   Pclass       418 non-null    int64  
 2   Name         418 non-null    object 
 3   Sex          418 non-null    object 
 4   Age          418 non-null    float64
 5   SibSp        418 non-null    int64  
 6   Parch        418 non-null    int64  
 7   Fare         418 non-null    float64
 8   Embarked     418 non-null    object 
dtypes: float64(2), int64(4), object(3)
memory usage: 29.5+ KB
```

3. describe()로 숫자형 특성의 요약 정보를 알려줌

```
In [7]: training.describe()
```

Out[7]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

=> 이 과정을 통해 데이터의 변수의 이름 및 특성을 확인 할 수 있음

데이터 탐색

4. Titanic 변수 탐색

Survived	Pclass	Name	Sex	Age	Sibsp	Parch	Ticket	Fare	Cabin	Embarked
1: 생존 0: 사망	1: 1등석 2: 2등석 3: 3등석	승객 이름	승객의 성별	승객의 나이	함께 탑승한 형제 또는 배우자의 수	함께 탑승한 부모 또는 자녀의 수	티켓 번호	티켓 요금	선실 번호	탑승한 항구

=> 탐색

1. Fare의 경우 사람마다 나타나는 차이가 무엇일까?
2. Name의 경우 정해진 형식이 없고 ""가 나타나는 경우가 존재
3. Sex의 경우 Int 형식으로 변환할 필요성이 존재
4. Ticket의 경우 문자와 숫자가 혼용되어 존재하고 있음
5. Cabin과 Age에서 결측치(NaN) 존재
6. Embarked에서 S, C, Q의 의미 파악
7. Age와 Fare의 숫자만 실질적인 의미를 내포하고 있음

결측치 처리

전처리&EDA

- 정의 및 과정
- 문제 정의
- 데이터 탐색
- 결측치 처리
- 상관관계 파악

캐글 (Kaggle)

1. 결측 데이터 종류

↓

2. 결측치 유형 탐색하기(표 만들기, 결측치 간 상관관계)

↓

3. 결측 데이터의 원인 및 각각의 원인에 따른 처리 방법론

↓

4. 결측치 처리 방법 선택(합리적 접근법)

↓

5. 결측치 처리 방법 선택(완전 제거법)

↓

5. 결측치 처리 방법 선택(다중대체)

Titanic - Machine Learning from Disaster

결측치 처리

1. 결측 데이터의 종류

- **완전 무작위 결측(MCAR : Missing completely at random)**

  변수 상에서 발생한 결측치가 다른 변수들과 아무런 상관이 없는 경우 우리는 **완전 무작위 결측(MCAR)**이라 함.
  대부분의 결측치 처리 패키지가 MCAR을 가정으로 하고 있고 일반적으로 우리가 생각하는 결측치를 의미함
  이러한 결측치는 보통 제거하거나 대규모 데이터 셋에서 단순 무작위 표본추출을 통해서 완벽한 데이터셋으로 만들어짐

  ex) 데이터를 입력하는 사람이 깜빡하고 입력을 안 했다든지 전산 오류로 누락된 경우 등

- **무작위 결측(MAR : Missing at random)**

  누락된 자료가 특정 변수와 관련되어 일어나지만, 그 변수의 결과는 관계가 없는 경우를 의미함.
  그리고 누락이 전체 정보가 있는 변수로 설명이 될 수 있음을 의미함.

  ex) 남성은 우울증 설문 조사에 기입 할 확률이 적지만 우울함의 정도와는 상관이 없는 경우

- **비 무작위 결측(MNAR : Missing at not random)**

  위의 두가지 유형이 아닌 경우를 MNAR이라고 함.
  MNAR은 누락된 값(변수의 결과)이 다른 변수와 연관 있는 경우를 의미함.

  ex) 남성이 우울증 설문 조사에 기입하는게 우울증의 정도와 관련이 있다면 이것은 MNAR임.

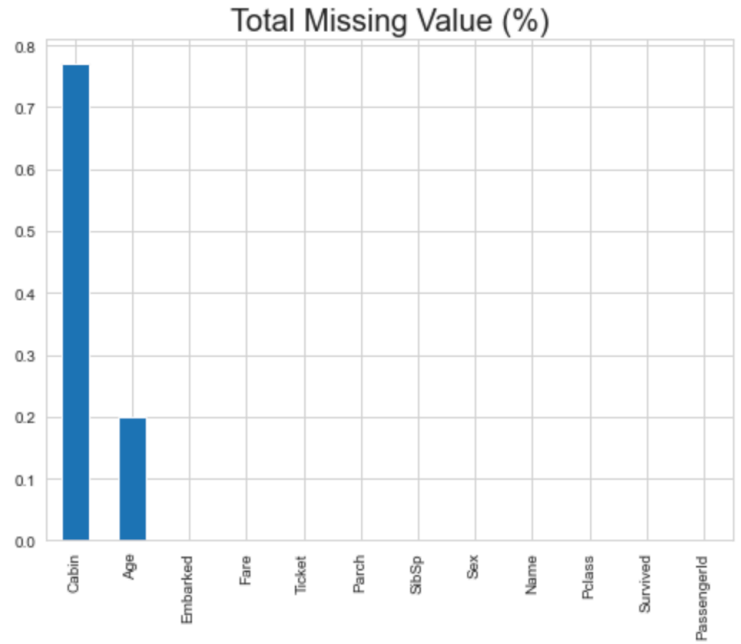
- https://en.wikipedia.org/wiki/Missing_data 참고

결측치 처리

2. 결측 값 유형 탐색하기

In [5]:
```
total = training.isnull().sum().sort_values(ascending=False)
percent = (training.isnull().sum()/training.isnull().count()).sort_values(ascending=False)
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
percent_data = percent.head(20)
percent_data.plot(kind="bar", figsize = (8,6), fontsize = 10)
plt.xlabel("", fontsize = 20)
plt.ylabel("", fontsize = 20)
plt.title("Total Missing Value (%)", fontsize = 20)
```

Out[5]:
```
Text(0.5, 1.0, 'Total Missing Value (%)')
```



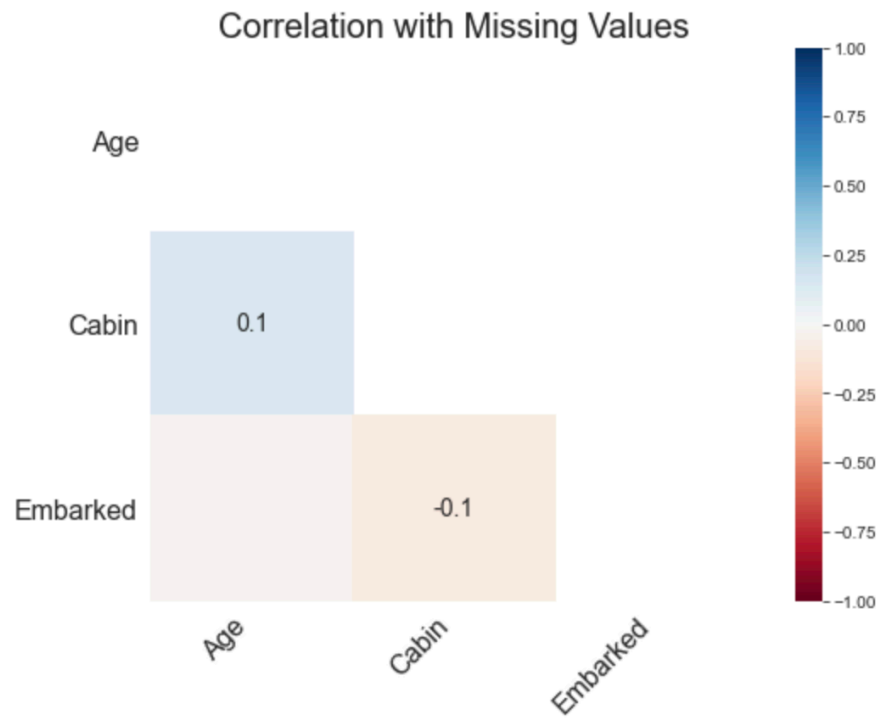
Cabin, Age, Embarked에서 결측치가 발생함을 알 수 있음

결측치 처리

2. 결측 값 유형 탐색하기

```
In [9]: import missingno as msno
        missingdata_df = training.columns[training.isnull().any()].tolist()
        msno.heatmap(training[missingdata_df], figsize=(8,6))
        plt.title("Correlation with Missing Values", fontsize = 20)
```

```
Out[9]: Text(0.5, 1.0, 'Correlation with Missing Values')
```



0.1과 -0.1로 보아 서로 상관이 없는 변수로 완전 무작위 결측(MCAR)임을 알 수 있다.

Titanic - Machine Learning from Disaster

결측치 처리

3. 결측 데이터의 원인 및 각각의 원인에 따른 처리 방법론

Cabin, Age, Embarked 순으로 77%, 19.8%, 0.002%로 Cabin은 생존과 관련이 없는 요소 제거하고 Age와 Fare를 대체

결측치 비율	처리방법
10% 미만	제거 또는 치환
10% 이상 20% 미만	모델 기반 처리
20% 이상	모델 기반 처리

- 제거

결측치가 발생한 행 또는 열을 삭제하는 단순한 방식

ex) pd.dropna(), pd.drop(), del name[column]

- 치환(합리적 접근)

결측치를 적당한 값으로 대체하는 방법

ex) 평균, 중앙값, 최빈 값으로 채울 수 있음

- 모델 기반 처리

결측치를 예측하는 새로운 모델을 구성하고 이를 채워나감

ex) Python KNN, 다중선형회귀 / R Mice 이용

결측치 처리

3. 결측 데이터의 원인 및 각각의 원인에 따른 처리 방법론

Cabin, Age, Embarked 순으로 77%, 19.8%, 0.002%로 Cabin은 생존과 관련이 없는 요소 제거하고 Age와 Fare를 대체

결측치 비율	처리방법
10% 미만	제거 또는 치환
10% 이상 20% 미만	모델 기반 처리
20% 이상	모델 기반 처리

- 제거

결측치가 발생한 행 또는 열을 삭제하는 단순한 방식

ex) pd.dropna(), pd.drop(), del name[column]

- 치환(합리적 접근)

결측치를 적당한 값으로 대체하는 방법

ex) 평균, 중앙값, 최빈 값으로 채울 수 있음

- 모델 기반 처리

결측치를 예측하는 새로운 모델을 구성하고 이를 채워나감

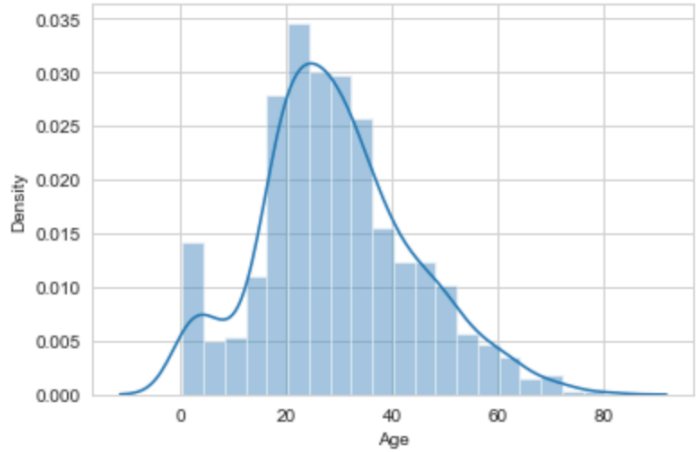
ex) Python KNN, 다중선형회귀 / R Mice 이용

결측치 처리

※ 치환(합리적 접근) 사용 시 주의

In [12]:
```
# copy 함수 같은 경우 파이썬 고유의 함수로 객체를 복사해서 옮기는 함수를 의미
copy = training.copy()
# nan 값을 없앰
copy.dropna(inplace = True)
# seaborn에서 distplot는 히스토그램과 커널 밀도 곡선을 나타내주는 함수를 의미
sns.distplot(copy["Age"])
```

Out[12]:
```
<AxesSubplot:xlabel='Age', ylabel='Density'>
```



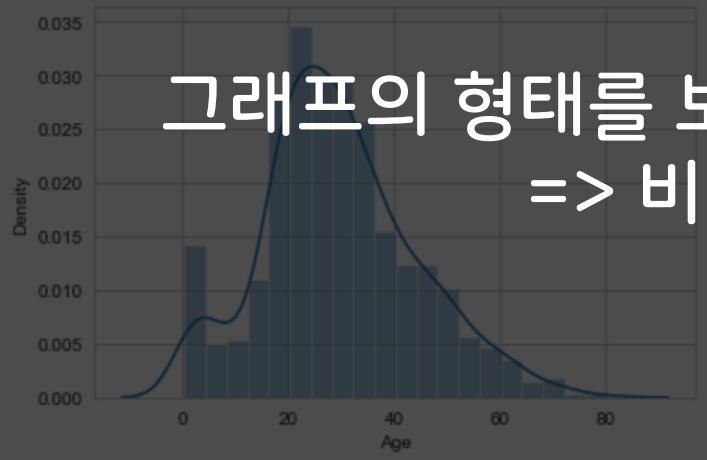
이런 그래프의 형태를 취해 줄 때 어떤 것으로 대체하는 것이 좋을까?

결측치 처리

※ 치환(합리적 접근) 사용 시 주의

```
In [12]: # copy 함수 같은 경우 파이썬 고유의 함수로 객체를 복사해서 옮기는 함수를 의미
         copy = training.copy()
         # nan 값을 없앰
         copy.dropna(inplace = True)
         # seaborn에서 distplot는 히스토그램과 커널 밀도 곡선을 나타내주는 함수를 의미
         sns.distplot(copy["Age"])

Out[12]: <AxesSubplot:xlabel='Age', ylabel='Density'>
```



그래프의 형태를 보고 결정할 수 있다
=> 비대칭도

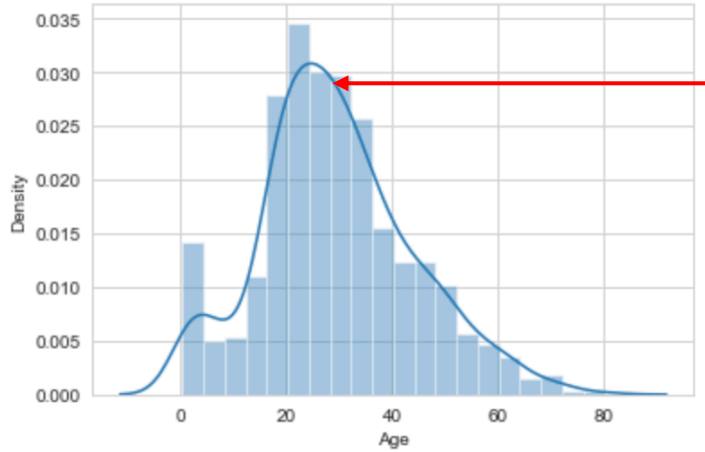
이런 그래프의 형태를 취해 줄 때 어떤 것으로 대체하는 것이 좋을까?

결측치 처리

※ 치환(합리적 접근) 사용 시 주의

```
In [12]: # copy 함수 같은 경우 파이썬 고유의 함수로 객체를 복사해서 옮기는 함수를 의미
copy = training.copy()
# nan 값을 없앰
copy.dropna(inplace = True)
# seaborn에서 distplot는 히스토그램과 커널 밀도 곡선을 나타내주는 함수를 의미
sns.distplot(copy["Age"])
```

```
Out[12]: <AxesSubplot:xlabel='Age', ylabel='Density'>
```



오른 쪽으로 치우쳐져 있으므로 평균 값으로 대체할 경우 왜도가 더욱 심화됨 따라서 중앙값으로 대체 해야함

https://ko.wikipedia.org/wiki/%EB%B9%84%EB%8C%80%EC%B9%AD%EB%8F%84 참고

상관관계 파악

※ 속성 간의 관계 분석하기

Categorical Variable (Qualitive)	Nomial Data	원칙은 숫자로 표현하면 안되나 편의상 숫자화 함 (순위의 개념은 없음) ex) 남자-0 여자-1
	Ordinal Data	원칙은 숫자로 표현하면 안되나 편의상 숫자화 함 (순위의 개념이 존재) ex) 소득 분위 10분위 > 9분위
Numeric Variable (Quantitative)	Continuous Data	데이터가 연속량으로 셀 수 있는 형태. ex) 키 - 166.1cm
	Discrete Data	데이터가 비 연속량으로 셀 수 있는 형태 ex) 자식 수 5명

데이터 조합	요약 통계	시각화
Categorical-Categorical	교차 테이블	모자이크 플롯
Numeric-Categorical	카테고리별 통계 값	박스 플롯
Numeric-Numeric	상관 계수	산점도

상관관계 파악

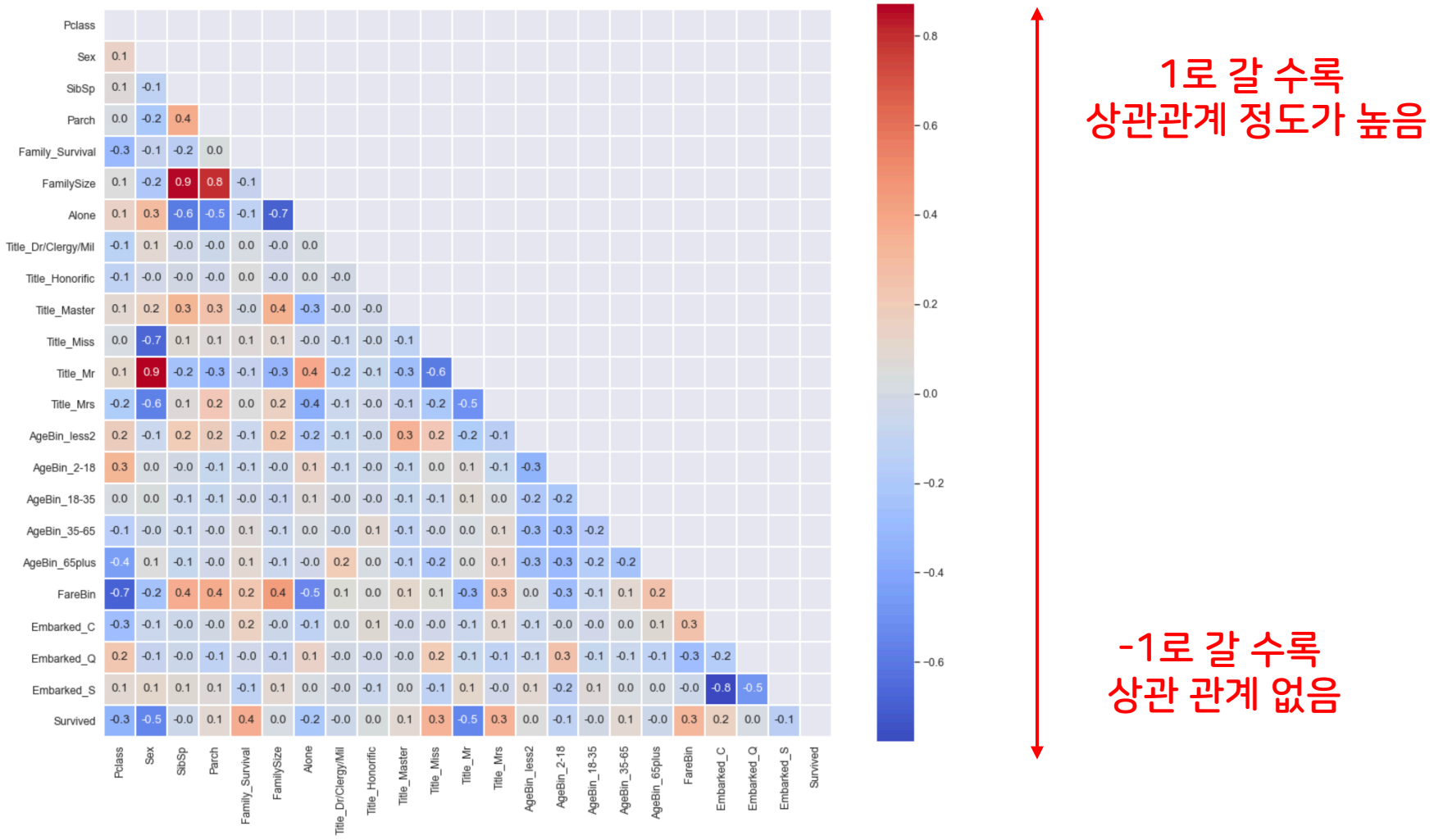
전처리&EDA

- 정의 및 과정
- 문제 정의
- 데이터 탐색
- 결측치 처리
- 상관관계 파악

캐글 (Kaggle)

※ 속성 간의 관계 분석하기

두 변수가 선형 또는 비선형적 관계를 가지고 있는지 살펴보는 것임



1로 갈 수록
상관관계 정도가 높음

-1로 갈 수록
상관 관계 없음

Titanic - Machine Learning from Disaster

캐글 (Kaggle)

전처리&EDA

- 정의 및 과정
- 문제 정의
- 데이터 탐색
- 결측치 처리
- 상관관계 파악

캐글 (Kaggle)

캐글러 들은 어떻게 EDA를 진행했을까?

https://github.com/CEOJINSUNG/Bigdata/blob/master/BigData%20Study/1%EC%A3%BC%EC%B0%A8/1%EC%A3%BC%EC%B0%A8_%EB%8D%B0%EC%9D%B4%ED%84%B0%20%EC%A0%84%EC%B2%98%EB%A6%AC%26EDA(%EB%B0%9C%ED%91%9C%EC%9A%A9).ipynb

Titanic - Machine Learning from Disaster

Check-Out

다음 주 발표자 자원 받습니다
그리고 다음주 공부 각오 및 계획을 알려주세요

발표자 2명 Check-In & Check-Out 질문 준비