스터디 2주차
데이터분석 과정

ABOUT MEETUP

2주차 스터디 내용

1 데이터에 대한 이해.

2 세부적인 데이터분석목적과 이에 따른 과정

3 데이터 전처리과정에 대해서 세부적으로 살펴봄.

4 실습 : 타이타닉 데이터분석

1. 데이터란?

5감(눈,귀,촉각 등)을 활용한 관측 -> 객관적이거나 주관적인 수치로 설명한 자료

통계학의 기원은 독일의 국가관리에서 출발

- 국가관리에서 가장 핵심은 "세금", 세금을 걷으려면 정확한 인구조사가 필수.
- Statistics는 국가라는 의미의 Status가 어원이고, Census는 라틴어 Censere 세금을 의미함.
- 즉 현재의 상태를 관측하고 이를 수치로 기록하는 것부터 출발

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S

1. 데이터란?

이러한 데이터는 행과, 열을 갖는 사각형구조를 말하며 이를 테이블이라 부름
도메인 마다 부르는 용어는 다르지만, 보통 **열을 변수**, **행을 관측치**라 부른다.

변수

관측값

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S

1. 데이터분석의 목적

데이터분석은 6가지 목적을 갖고 목적에 따라 과정이 달라진다.

(1) 현상을 설명 : 기술통계분석(Descriptive analysis)

- 데이터를 통해 **“지금의 상태를 설명(현상)을”** 설명 하는것이 목표
- 생존자는 얼마나 될까?, 탑승객의 나이는 어땠을까?

(2) 여러 데이터간 관계를 탐색: 탐색적 데이터분석(Exploratory Data Analysis)

- 생존과 강한 연관이 있는 요인들은 어떤 요인일까?
- 가설이라는것이 등장.

(3-1) 내가 살펴본 데이터와 결과가 일반적인 현상일까?: 추론적 데이터분석(Inferential Data Analysis)

- P-Value, 통계적 검증
- 여러 사고 데이터를 통해 일반적인 사고에서 나이가 진짜로 영향을 미치는지 살펴본다.

1. 데이터분석의 목적

데이터분석은 6가지 목적을 갖고 목적에 따라 과정이 달라진다.

(4) 앞으로는 어떻게 될까? : 예측분석(Predictive analysis)

- 내가 갖은 데이터를 통해 앞으로의 일을 예측하자.
- 로지스틱 회귀분석을 통한 생존자 예측.

(5) 맞는지 틀린지 진짜로 실험 해보자:인과분석(Causal analysis)

- 데이터간 인과관계를 실험을 통해 살펴본다
- A/B테스트, AA 테스트

(6) 법칙을 만든다: 매커니즘 분석 (Mechanistic analysis)

- E = mc^2 , 작용반작용의 법칙 등

1. 데이터분석과정

데이터분석은 목적에 따라 3가지 과정을 갖는다.

(1) 현상을 설명 : 기술통계분석(Descriptive analysis)

- 데이터를 통해 **“지금의 상태를 설명(현상)을”** 설명 하는것이 목표
- 생존자는 얼마나 될까?, 탑승객의 나이는 어땠을까?

(2) 여러 데이터간 관계를 탐색: 탐색적 데이터분석(Exploratory Data Analysis)

- 생존과 강한 연관이 있는 요인들은 어떤 요인일까?
- 가설이라는것이 등장.

(3) 내가 살펴본 데이터와 결과가 일반적인 현상일까?: 추론적 데이터분석(Inferential Data Analysis)

- P-Value, 통계적 검증
- 여러 사고 데이터를 통해 일반적인 사고에서 나이가 진짜로 영향을 미치는지 살펴본다.

1. 데이터분석의 목적

데이터분석은 6가지 목적을 갖고 목적에 따라 과정이 달라진다.

(3-1) 앞으로는 어떻게 될까? : 예측분석(Predictive analysis)

- 내가 갖은 데이터를 통해 앞으로의 일을 예측하자.
- 로지스틱 회귀분석을 통한 생존자 예측.

(3-1) 맞는지 틀린지 진짜로 실험 해보자:인과분석(Causal analysis)

- 데이터간 인과관계를 실험을 통해 살펴본다
- A/B테스트, AA 테스트

1. 기술분석 : 지금 상황은 어떤가?

1.목적 :

- 데이터를 통해 상황을 파악한다.
- 데이터 자체를 **이해한다.**

2. 왜 할까?

- 테이블은 읽기 힘들다.
- 한눈에 파악하기 위함.

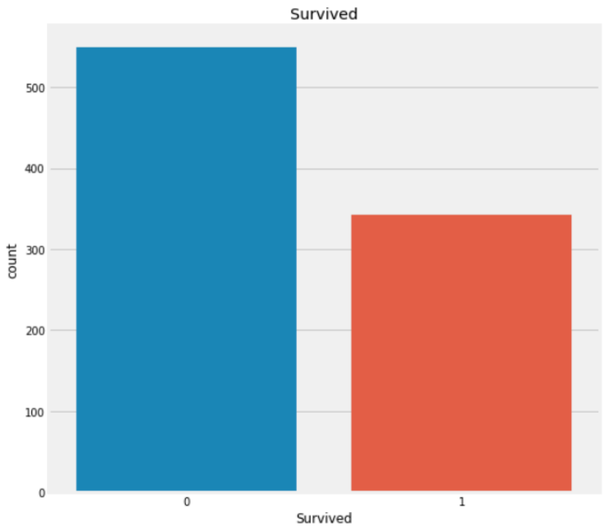
가장 오래된 분석방법으로 통계학의 기원으로 부터 시작하였음.

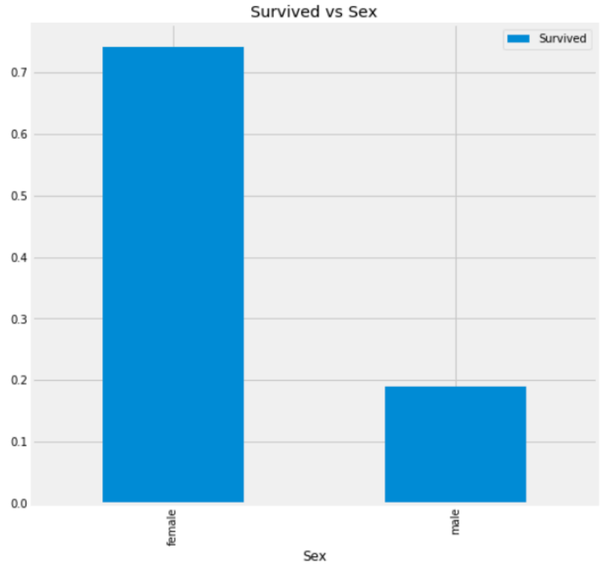
통계학의 기원은 독일의 국가관리에서 출발

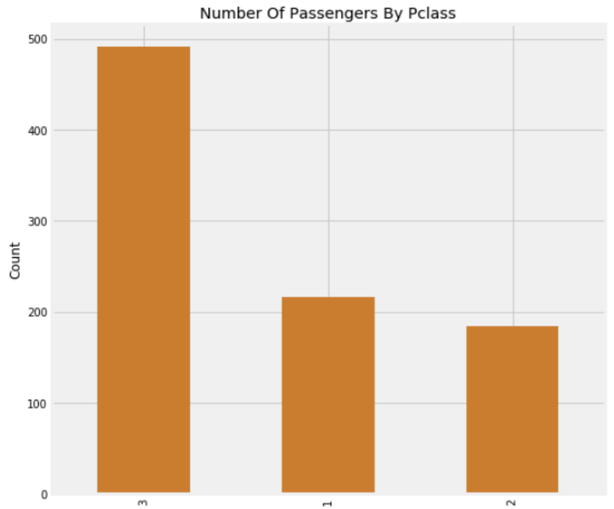
- 국가관리에서 가장 핵심은 “세금”, 세금을 걷으려면 정확한 인구조사가 필수.
- Statistics는 국가라는 의미의 Status가 어원이고, Census는 라틴어 Censere 세금을 의미함.
- 즉 현재의 상태를 관측하고 이를 수치로 기록하는 것부터 출발

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S

01주차 :데이터분석의 과정







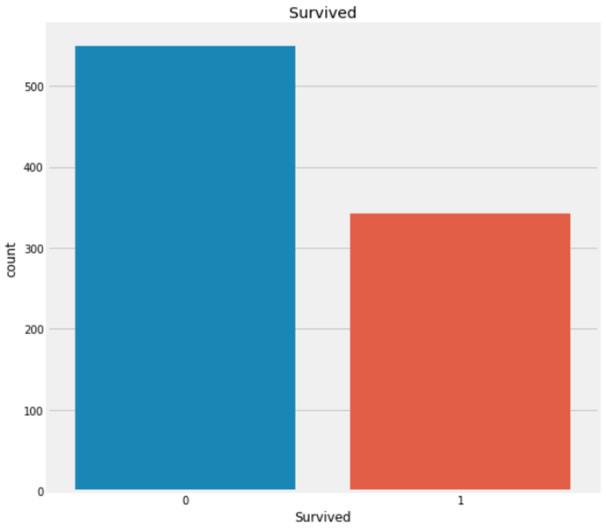
2. 여러 데이터간 관계를 탐색: 탐색적 데이터분석(Exploratory Data Analysis)

1.목적 :

- 데이터를 목적에 맞게 살펴본다.
- 목적과 데이터간의 관계(상관관계)를 파악한다.

2. 왜 할까?

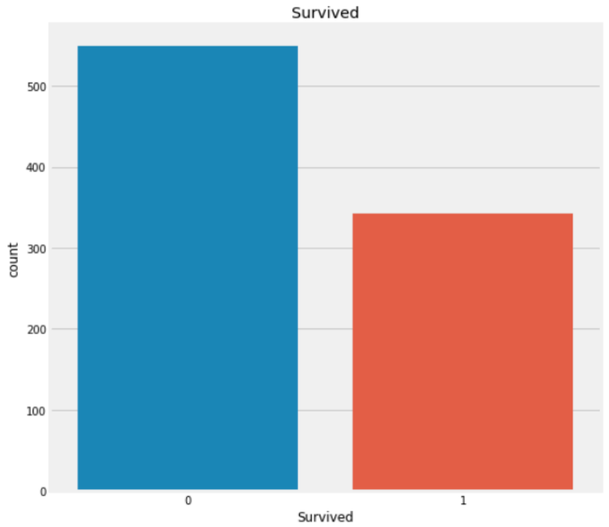
- 데이터 분석은 **인사이트**를 찾기 위함.
- 데이터를 통해 **목적을 달성**하기 시작하는 첫 단계
- 데이터를 통해 **다음 단계로의 힌트**를 얻는다.(**가설발견**)

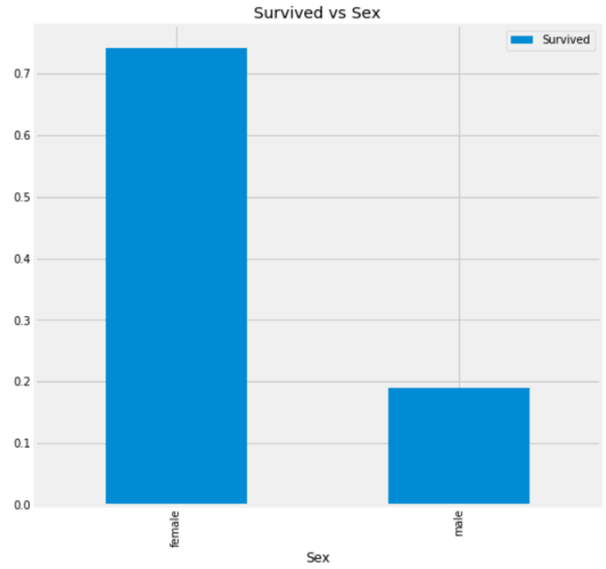


일단 사망자가 생존자 보다 월등하게 많다는 것을 알게됨.

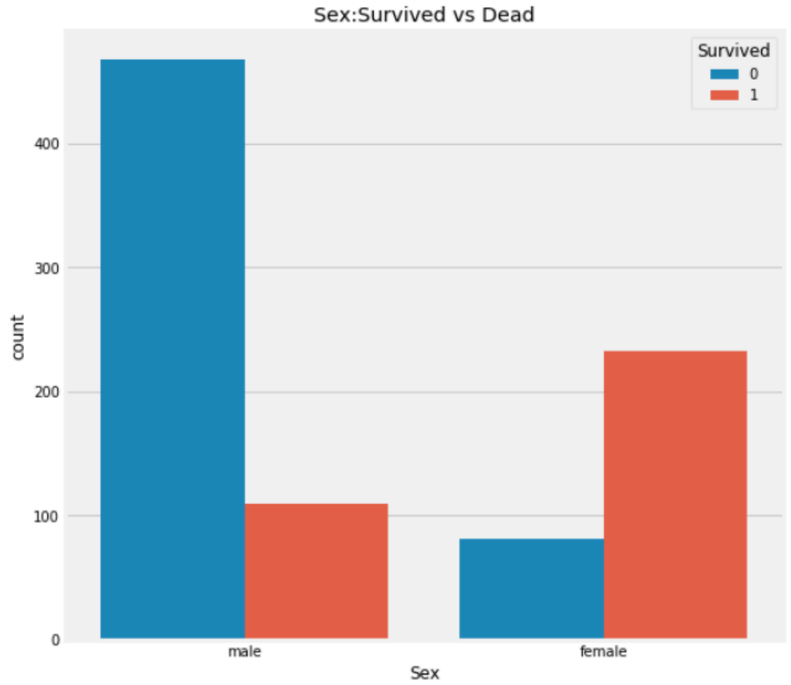
왜지?

생존과 사망에 영향을 끼친 요인들이 있지 않을까?

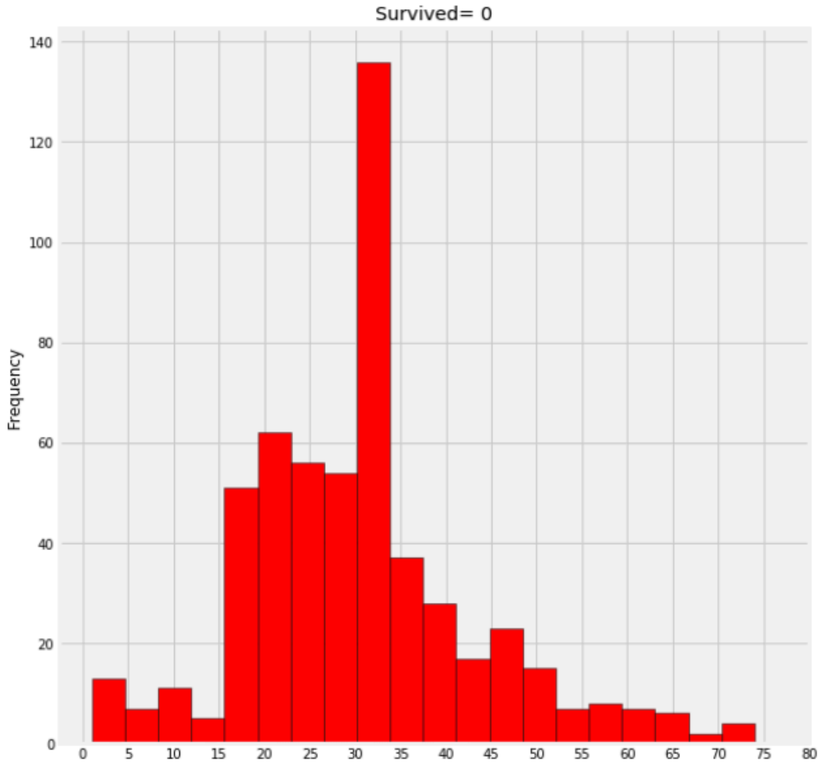


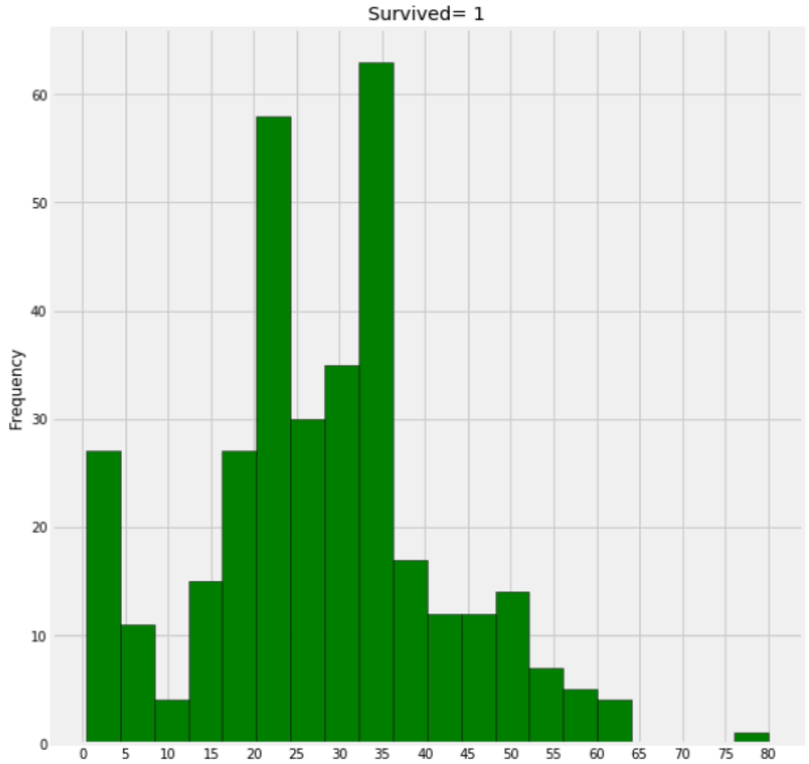


성별 생존을 살펴보았더니 여성이 남성에 비해서 많이 생존했네?

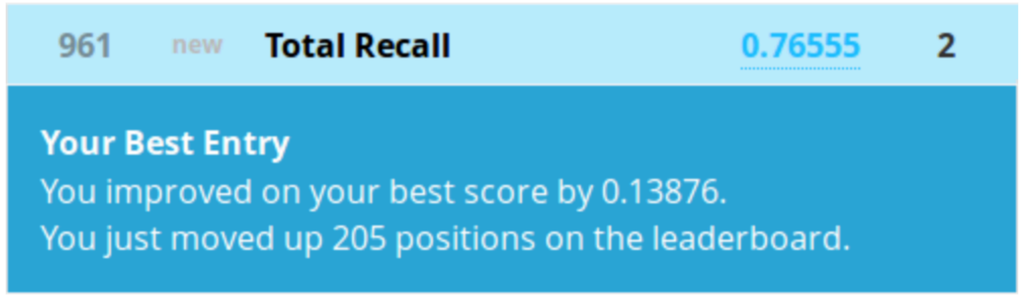


나이가 어릴수록 많이 생존했네?





3. 앞으로는 어떻게 될까? : 예측분석(Predictive analysis)



Q&A

실습 시작

그리고 과제