고급 소프트웨어 실습(CSE4152)

2주차

Exploratory Data Analysis (탐색적 데이터 분석, EDA)

목차

- 실험 환경
- Exploratory Data Analysis
- Pearson Correlation Coefficient
- 실습
- 과제

실험환경

- Python
- Jupyter Notebook

Exploratory Data Analysis

1. 정의

- 수집한 데이터가 들어왔을 때, 이를 다양한 각도에서 관찰하고 이해하는 과정이다. 즉, 데이터를 분석하기 전에 그래프나 통계적인 방법으로 자료를 직관적으로 바라보는 과정이다.

Exploratory Data Analysis

2. 필요한 이유

- 데이터의 분포 및 값을 검토함으로써 데이터가 표현하는 현상을 더 잘 이해하고, 데이터에 대한 잠재적인 문제를 발견할 수 있습니다. 이를 통해 본격적인 분석에 들어가기 앞서 데이터의 수집을 결정할 수 있습니다.
- 다양한 각도에서 살펴보는 과정을 통해 문제 정의 단계에서 미쳐 발생하지 못했을 다양한 패턴을 발견하고, 이를 바탕으로 기존의 가설을 수정하거나 새로운 가설을 세울 수 있습니다.

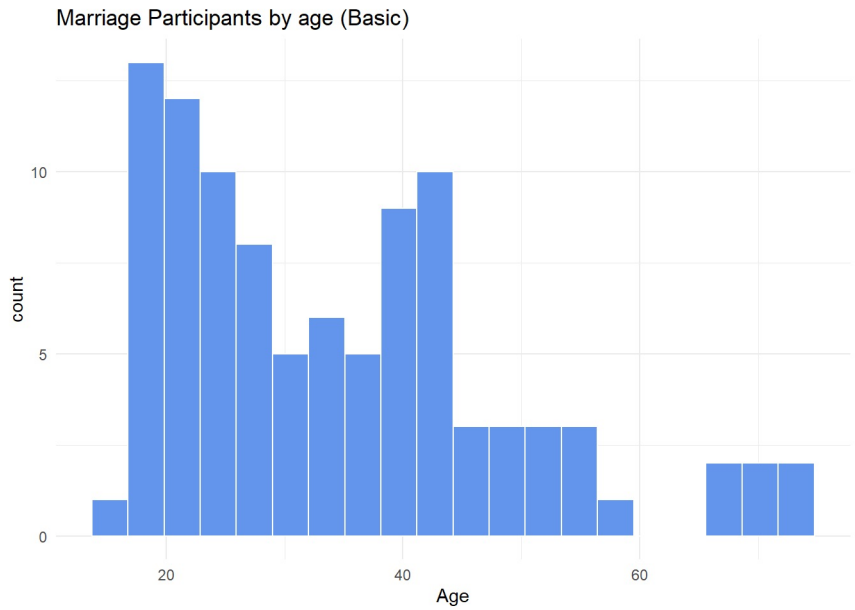
Exploratory Data Analysis

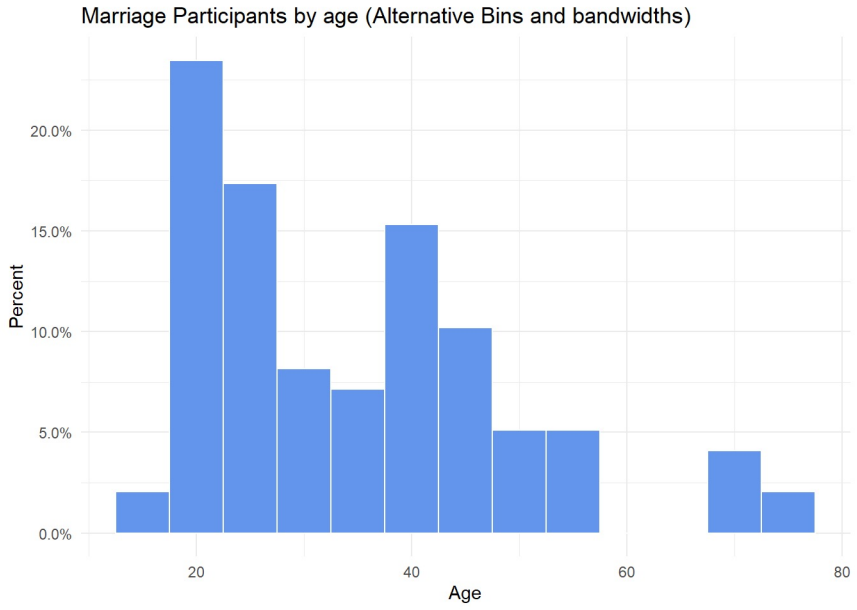
3. 과정

1. 분석의 목적과 변수가 무엇이 있는지 확인. 개별 변수의 이름이나 설명을 가지는지 확인
2. 데이터를 전체적으로 살펴보기
3. 데이터의 개별 속성값을 관찰
4. 속성 간의 관계에 초점을 맞추어 개별 속성 관찰에서 찾아내지 못했던 패턴을 발견 (상관관계, 시각화)

Exploratory Data Analysis

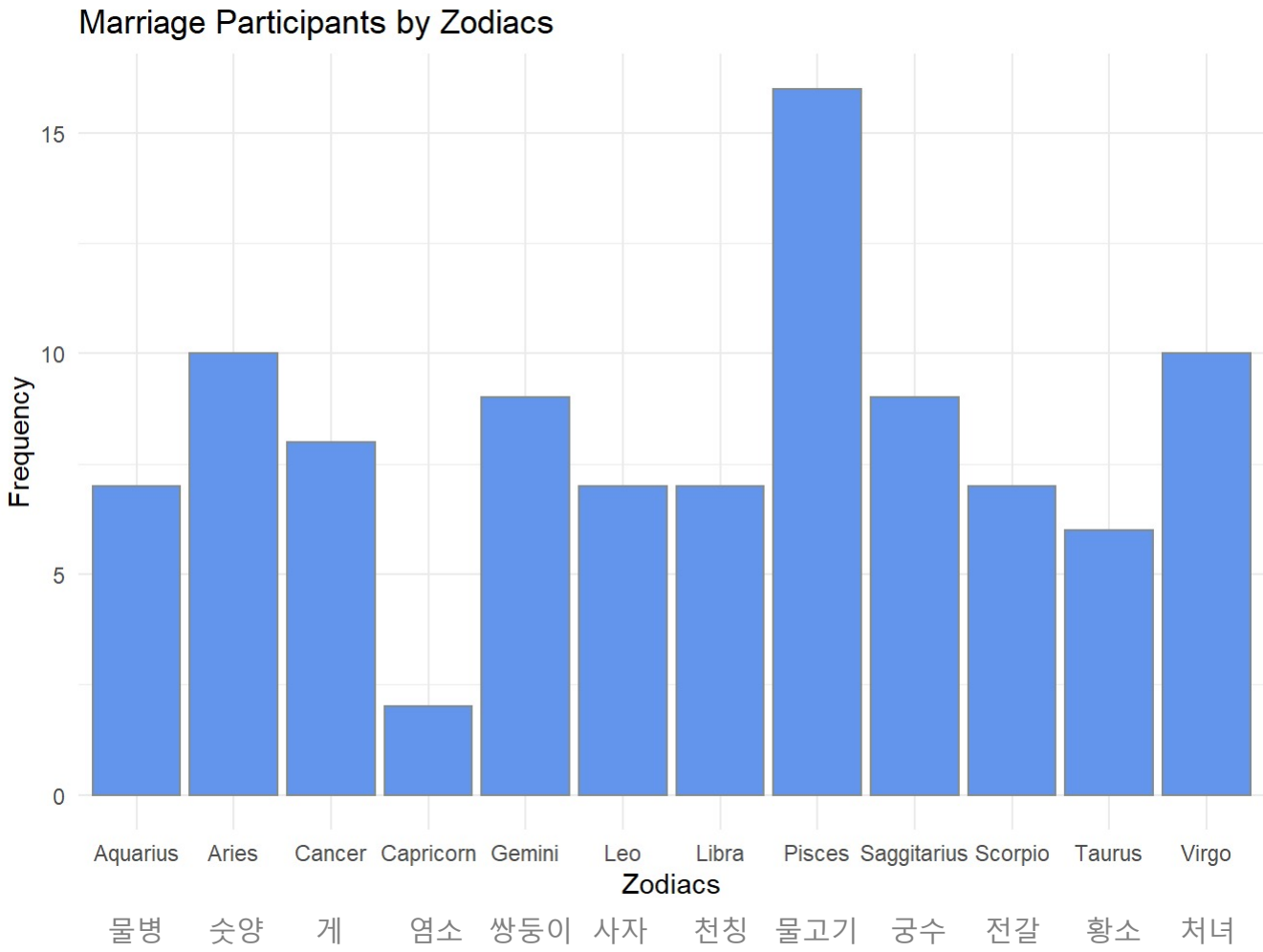
Histogram : distribution of a continuous variable





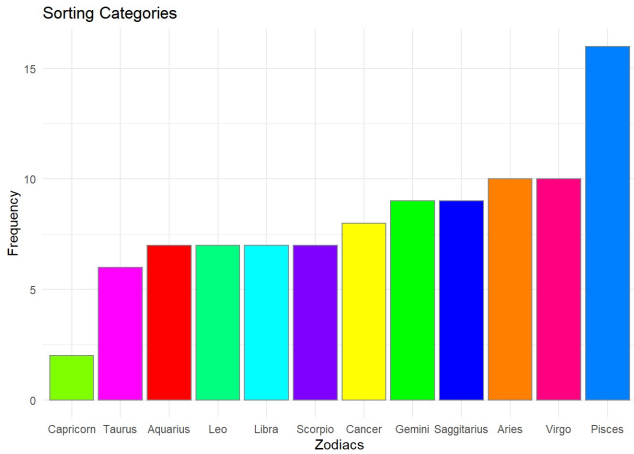
Exploratory Data Analysis

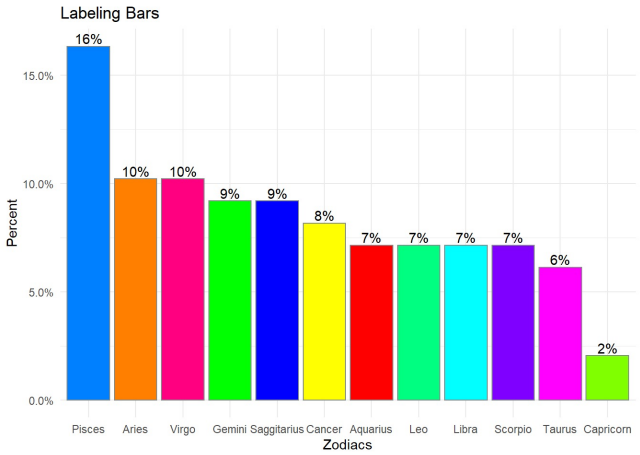
Bar chart : distribution of a categorical variable

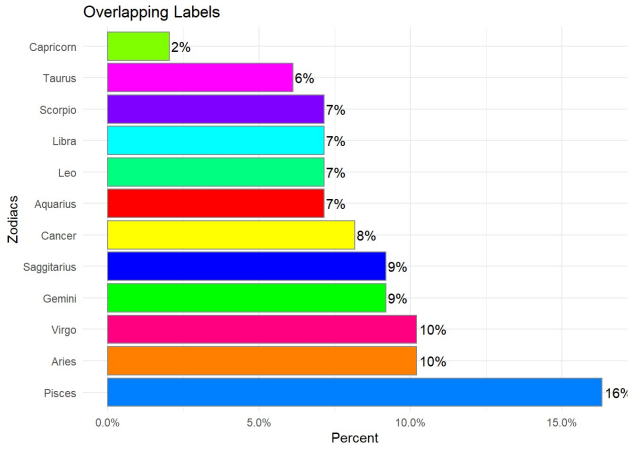


Exploratory Data Analysis

Bar chart

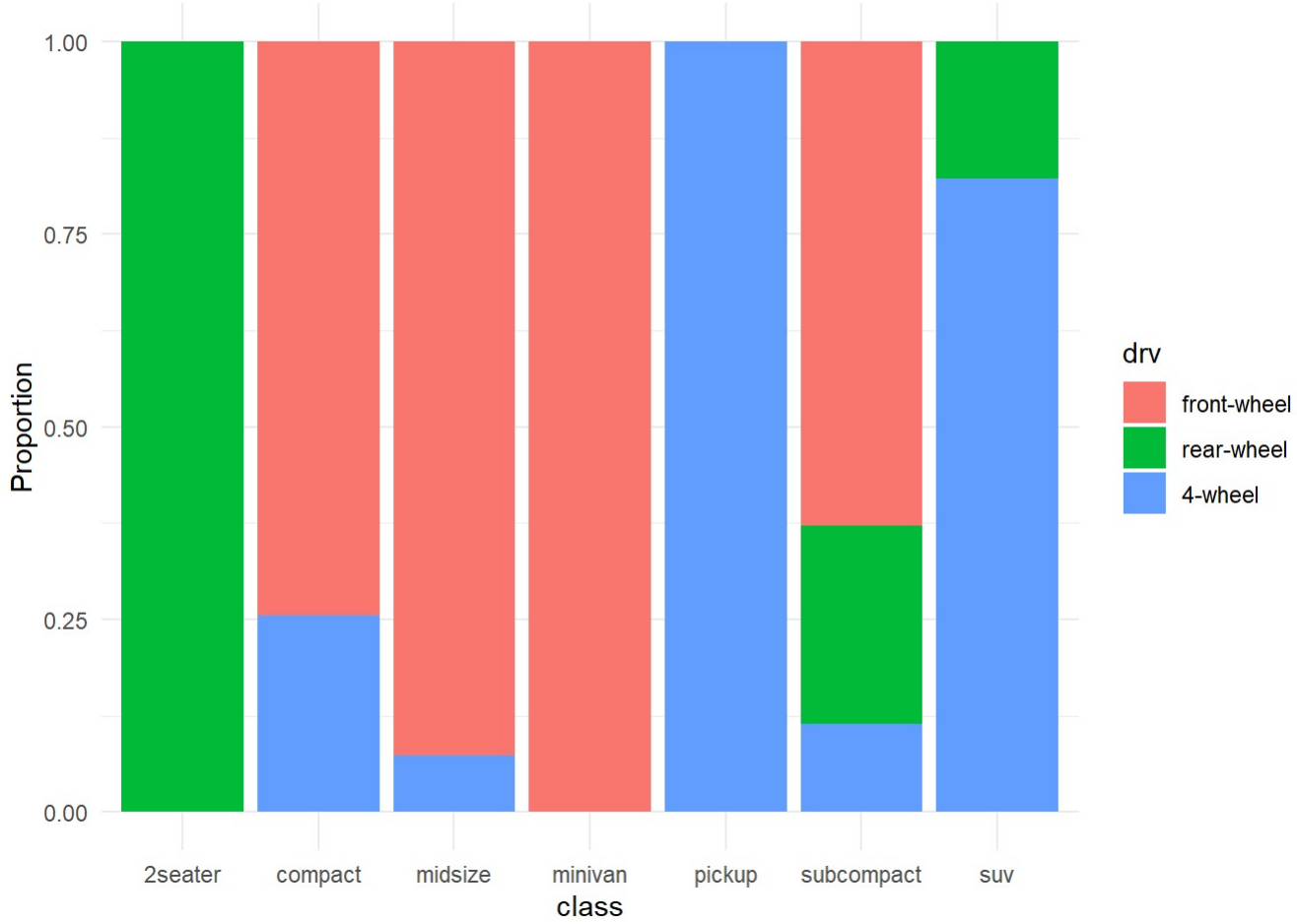






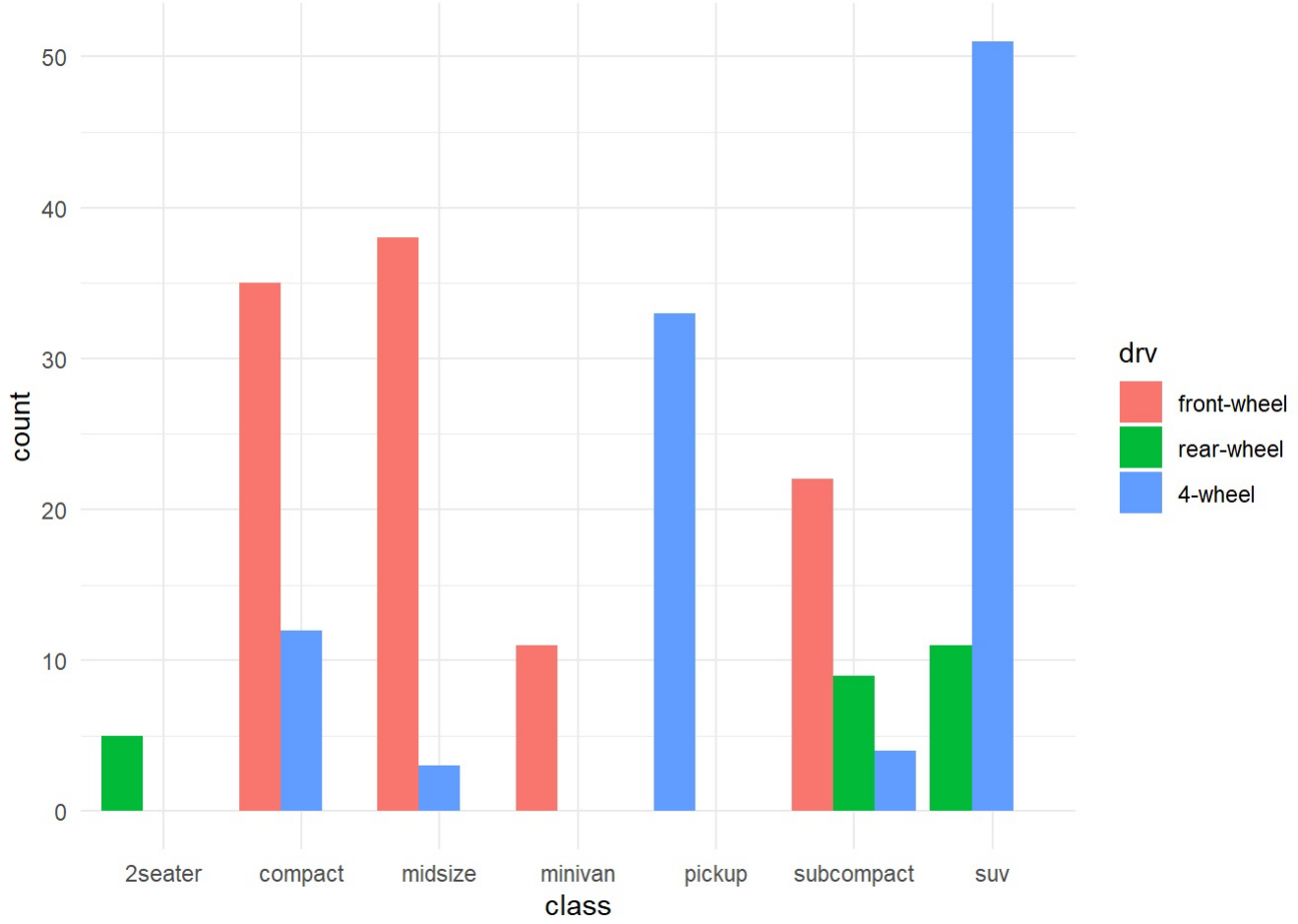
Exploratory Data Analysis

Stacked Bar Chart



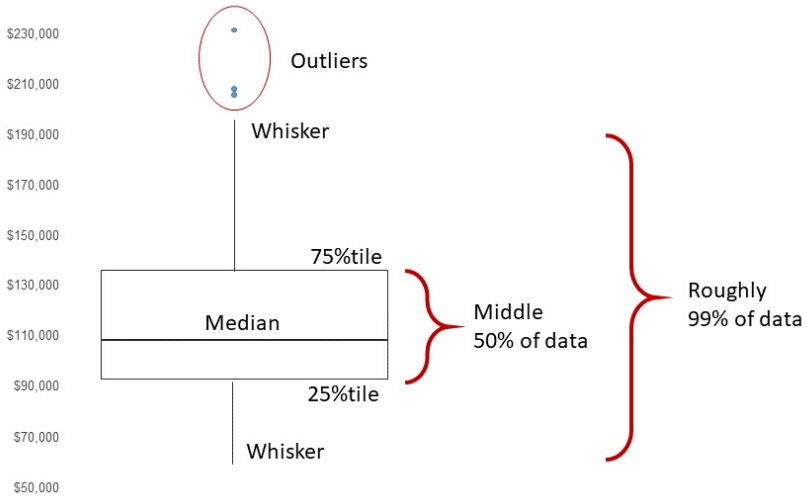
Exploratory Data Analysis

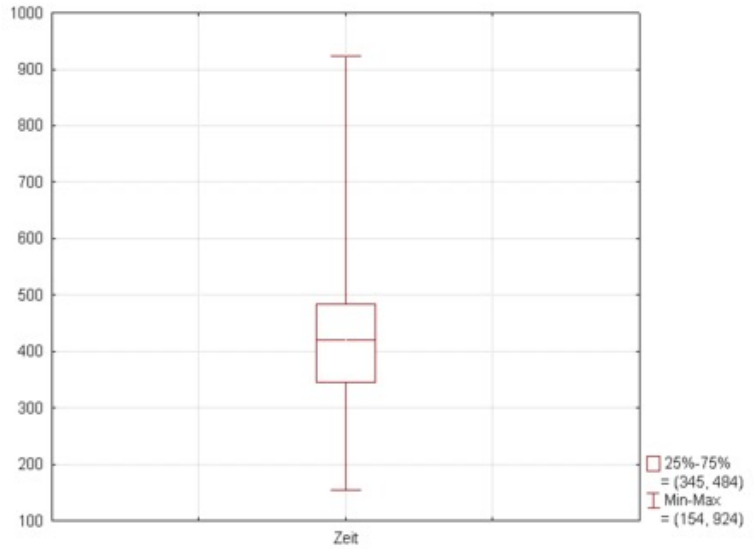
Grouped Bar Chart



Exploratory Data Analysis

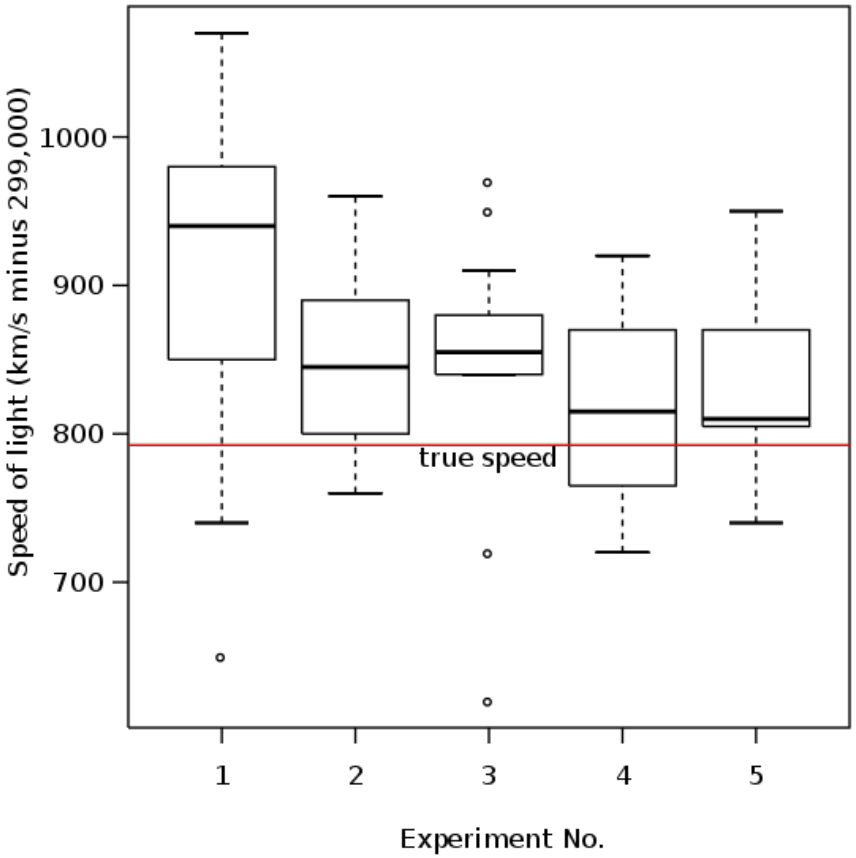
Boxplot





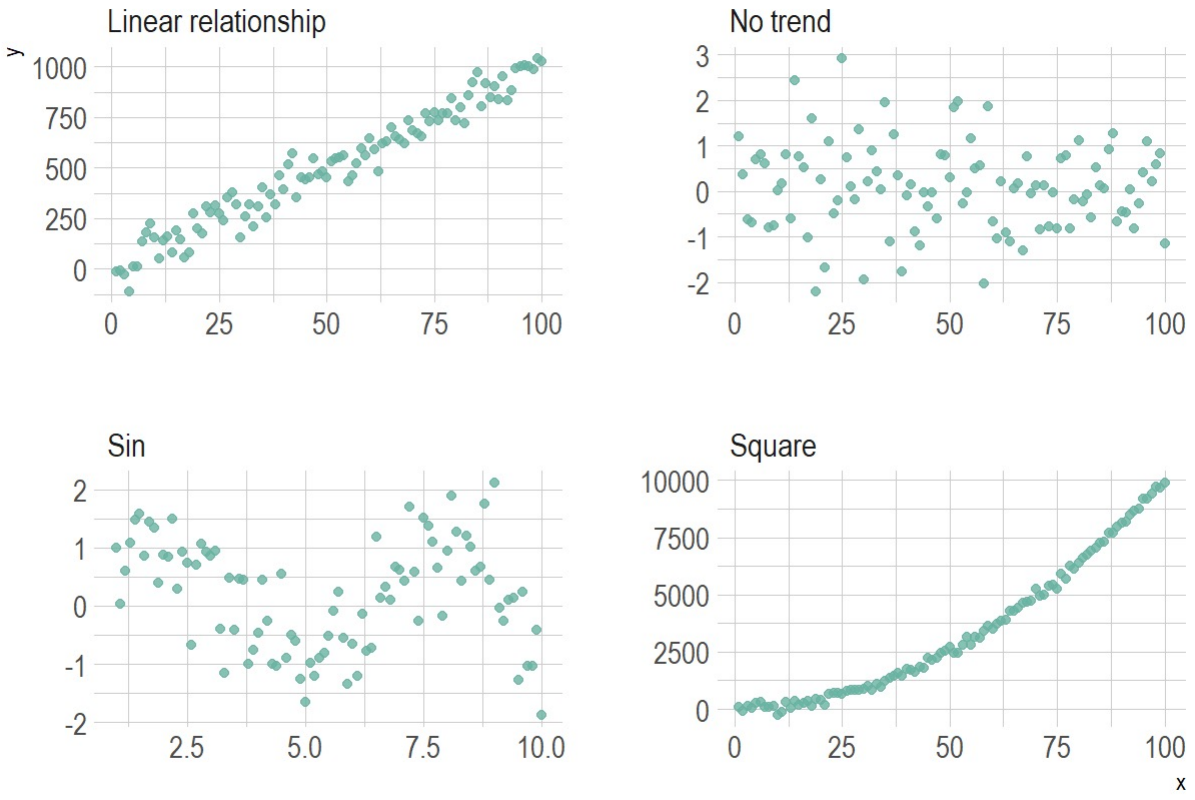
Exploratory Data Analysis

Boxplot



Exploratory Data Analysis

Scatter plot



Pearson Correlation Coefficient

- 통계학에서 피어슨 상관 계수란 두 변수 X와 Y 간의 선형 상관 관계를 계량화한 수치이다. 피어슨 상관 계수는 코시-슈바르츠 부등식에 의해 +1과 -1 사이의 값을 가지며, +1은 완벽한 양의 선형 상관 관계, 0은 선형 상관 관계 없음, -1은 완벽한 음의 선형 상관 관계를 의미한다.

$$r_{XY} = \frac{\frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{n}}{\sqrt{\frac{\sum_i^n (X_i - \bar{X})^2}{n}}\sqrt{\frac{\sum_i^n (Y_i - \bar{Y})^2}{n}}} \qquad r_{XY} = \frac{\frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}}{\sqrt{\frac{\sum_i^n (X_i - \bar{X})^2}{n-1}}\sqrt{\frac{\sum_i^n (Y_i - \bar{Y})^2}{n-1}}}$$

Pearson Correlation Coefficient

- SaratogaHouses Dataset: 집값 예측 데이터셋
  - Price: price
  - lotSize: size of lot
  - Age: age of house
  - landValue: value of land
  - livingArea: living area (square feet)
  - pctCollege: percent of neighborhood that graduated college
  - Bedrooms: number of bedrooms
  - Fireplaces: number of fireplaces
  - Bathrooms: number of bathrooms (half bathrooms have no shower or tub)
  - Rooms: number of rooms
  - Heating: type of heating system
  - Fuel: fuel used for heating
  - Sewer: type of sewer system
  - Waterfront: whether property includes waterfront
  - newConstruction: whether the property is a new construction
  - centralAir: whether the house has central air

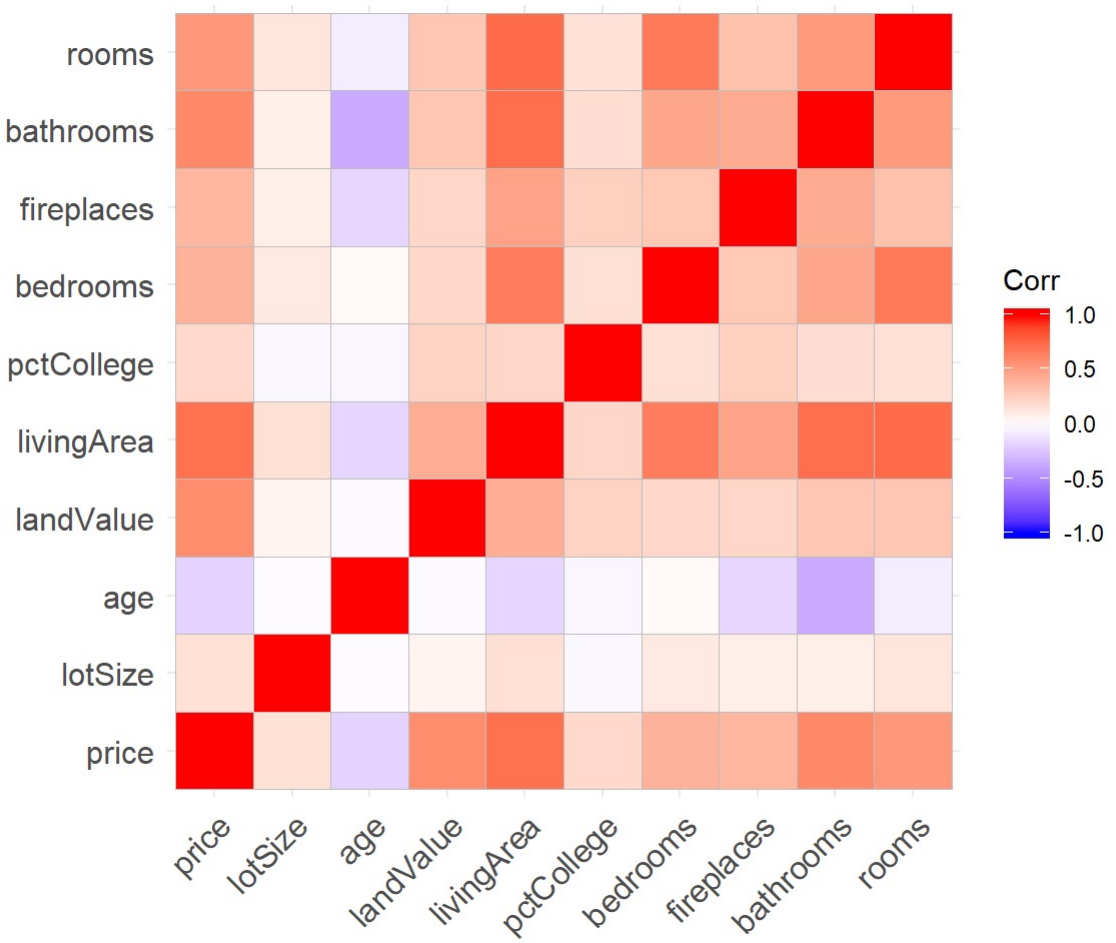
Pearson Correlation Coefficient

- 시각화 예시

Price
lotSize
Age
landValue
livingArea
pctCollege
Bedrooms
Fireplaces
Bathrooms
Rooms

숫자 데이터만 사용

Heating
Fuel
Sewer
Waterfront
newConstruction
centralAir



Dataset: Titanic

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

실습1. Pearson Correlation 함수

def pearsonCorrelation(data, source_column, target_column):

데이터가 주어지고 source_column, target_column 정보가 주어졌을때 pearson correlation 결과를 return 해주는 함수를 작성하시오.

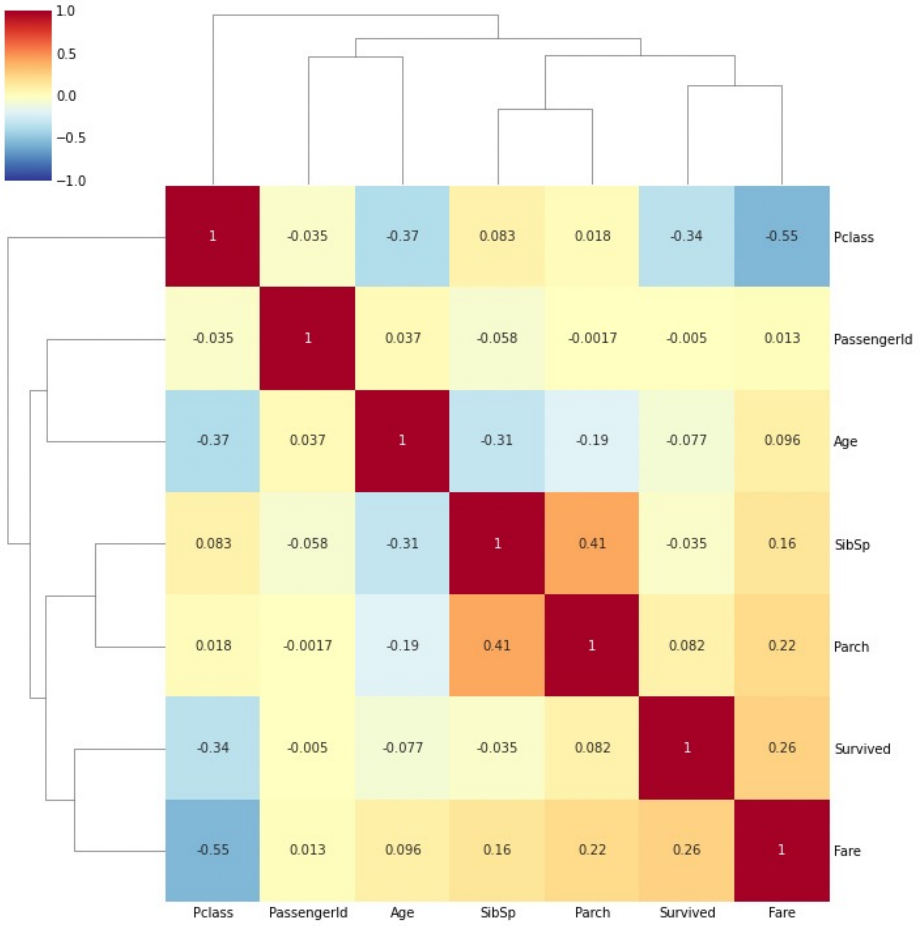
실습2. Pearson Correlation 함수 검증

작성된 pearsonCorrelation 함수가 주어진 데이터에서 작동하는지 확인한다. 작동이 안된다면 왜 안되는지, 확인하고 적절한 조치를 취한다.

실습3. Visualizing

정제된 correlation table 을 visualizing 한다.

예시:



실습4: 생존율에 영향을 미치는 요인 분석

- 이전 실습에서 확인한 Heatmap 을 확인하여 survived 에 영향을 주는 column 을 분석하고 이를 그림, 그래프, 도표로 표시한다.
- 4가지의 상관관계를 분석내용을 첨부하여 제출하시오.

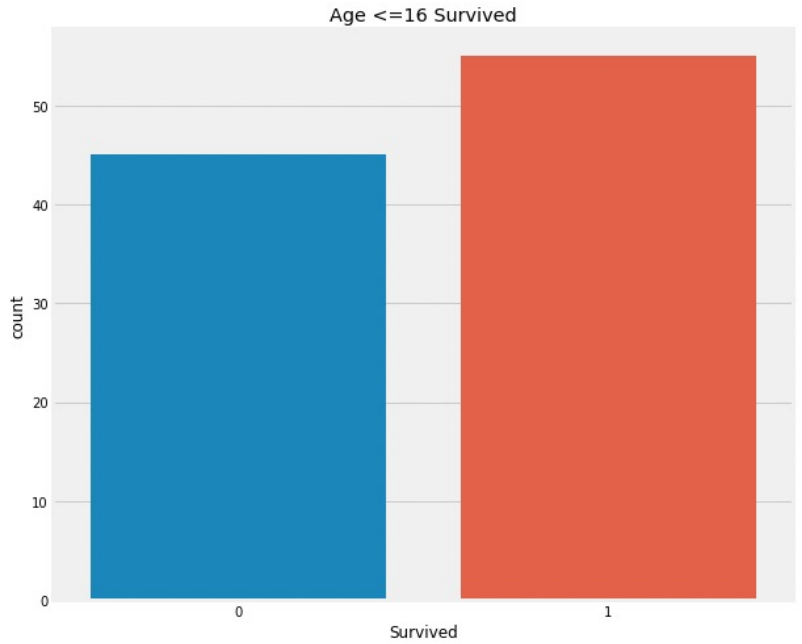
실습4. 결과 예시

아래의 예시들은 나이와 생존의 관계이다.

16세 이하의 생존율은 55% 로 다른 나이에 비해 생존율이 높다.

In [29]:
```
# 16세 이하의 생존률
dropped_data[(data['Age'] <= 16) & (data['Survived'] == 1)]['Survived'].count() / dropped_data[data['Age'] <= 16]['S
data1 = dropped_data[(data['Age'] <= 16)]

f,ax=plt.subplots(1, 1,figsize=(9,8))
sns.countplot('Survived',data=data1,ax=ax)
ax.set_title('Age <=16 Survived')
plt.show()
print(f"16세 이하의 생존율 {dropped_data[(data['Age'] <= 16) & (data['Survived'] == 1)]['Survived'].count() / dropped_da
```




```
16세 이하의 생존율 55.00000000000001%
```

과제1: Pearson Correlation Coefficient 함수

Pearson Correlation Coefficient 함수

$$r_{XY} = \frac{\frac{\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})}{n}}{\sqrt{\frac{\sum_i^n (X_i - \bar{X})^2}{n}}\sqrt{\frac{\sum_i^n (Y_i - \bar{Y})^2}{n}}}$$

가 아래의 수식과 동일한 표현이라는 것을 보이시오.

$$r_{XY} = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$