하루 동안 경험하는 데이터 분석

Decision Insight

시작하기 전에...

- 데이터 분석 전체의 흐름을 파악한다. (데이터 분석가의 업무 범위)
  데이터 전처리 -> 시각화 -> 모델링
- 전체를 이해하는 과정 중에 특정 과제는 직접 다루어 본다.
  데이터 전처리 부분 외 feature engineering에 해당하는 변수 변환, 특정 변수 제거, 신규 변수 도출 등을 다룬다.
- 탐색적 시각화를 통하여 tabular data로 알 수 없었던 숨어 있는 사실을 찾아내고 의사결정에 중요한 내용을 보고서 형태로 만드는 과정을 경험한다.

1. 목적과 필요성

- 목적

  하루 동안 데이터 분석의 처음과 끝을 경험한다.

  사용 데이터 : Titanic 데이터

- 필요성

  데이터 분석을 이루는 3개 기둥(데이터 전처리, 시각화, 모델링)을 개별적으로 학습하지 않고 하루에 전체 프로세스를 경험함으로써 실무 적용 및 활용도를 높인다.

2. 시간표

시간	제목	주요내용
1교시	R 설치	R과 RSTUDIO 설치, 주요 패키지 설치, 기 설치된 경우 패스
2교시	데이터 읽기	Titanic 데이터 읽기 및 데이터 전처리(결측치, 중복치, 이상치 등)
3교시	데이터 시각화 1	데이터 이해를 위한 데이터 시각화 1
4교시	데이터 시각화 2	데이터 이해를 위한 데이터 시각화 2
5교시	Feature Engineering	변수 선택 및 도출 변수 생성, 변수 변환
6교시	데이터 모델링 1	4개 분류 모형 선택 및 scripting
7교시	데이터 모델링 2	4개 분류 모형 튜닝 및 성능 비교
8교시	Publishing(배포)	웹 보고서 작성

3. 데이터 읽기

- R을 사용하는데 첫번째 장애물이 데이터 읽기입니다.
  1가지 문제를 다양한 방법으로 해결할 수 있다는 장점이 처음 시작하는 사용자에게
  오히려 방해 요인이 될 수 있습니다.

- 해결 방안
  rio 패키지의 import / export 함수만 기억하세요.

- 사용법
  엑셀파일이나 csv 파일을 읽어올 때 : import 함수 사용
  R 오브젝트(일종의 R 파일)를 엑셀파일이나 csv 파일로 저장할 때 : export 함수 사용

- 전처리 : 결측 데이터, 중복 데이터, 이상치 확인 및 처리

4.1 데이터 시각화

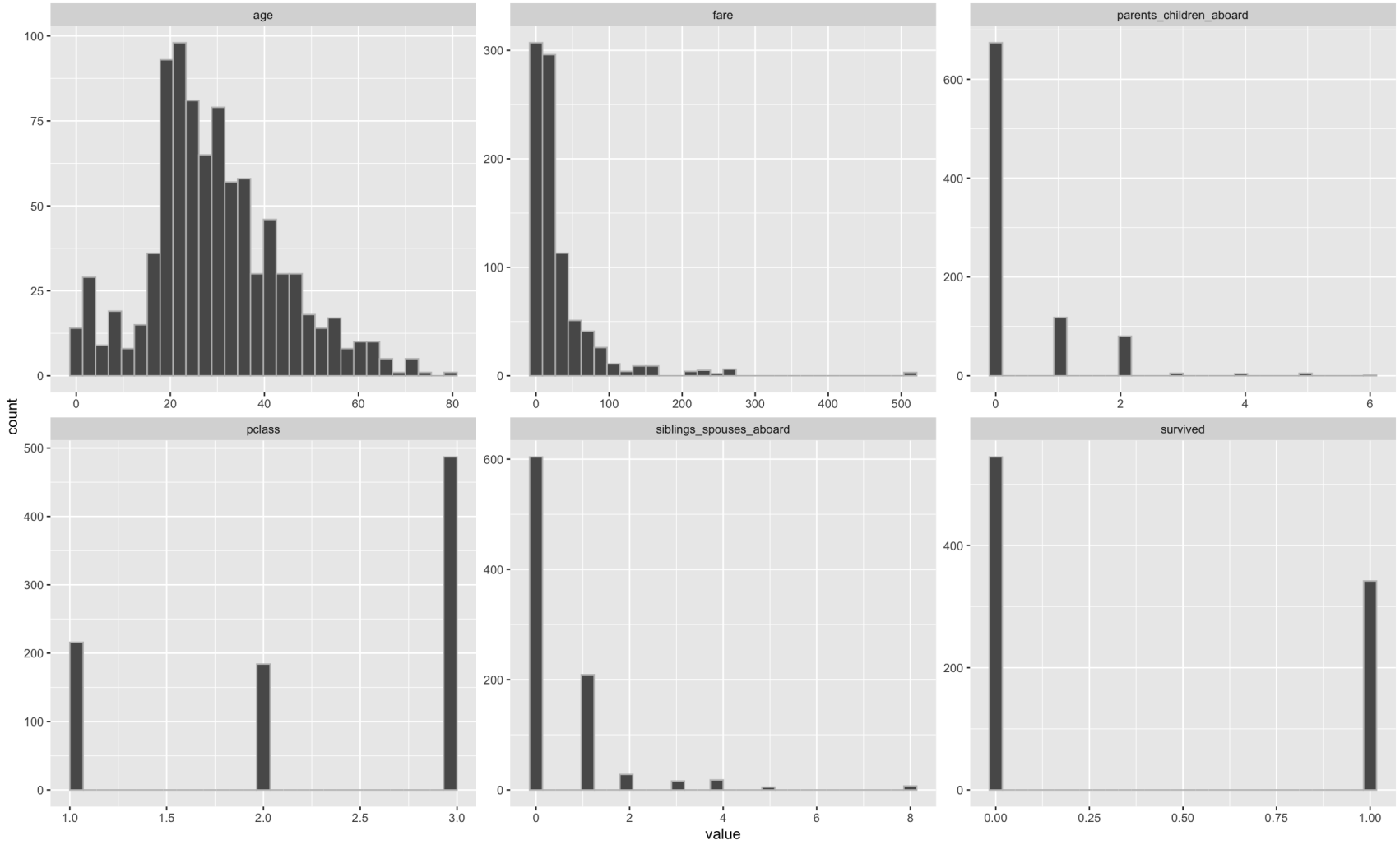
- 데이터 시각화는 2가지로 구분할 수 있습니다.

- 탐색적 시각화
  인간은 단 10줄의 데이터를 이해하는데 어려움을 느낍니다. 예를 들어 엑셀 시트에 A열에는 날짜, B열에는 매출이 각각 10개씩 있는 워크시트가 있다고 했을 때 이것을 한눈에 이해하기 어렵고 천천히 살펴봐야 하는 존재입니다. 그런데 조직 내 데이터는 수천, 수만 데이터로 쌓여있습니다. 데이터가 말하고자 하는 바를 이해하기는 쉽지 않습니다. 이러한 어려움을 해결할 수 있는 방법이 시각화입니다. 숫자보다는 그림을 봄으로써 데이터가 어떤 모습이며 무엇을 말하고 있는 지를 알 수 있는 장치인 것입니다.

- 실습 : ggplot2 패키지

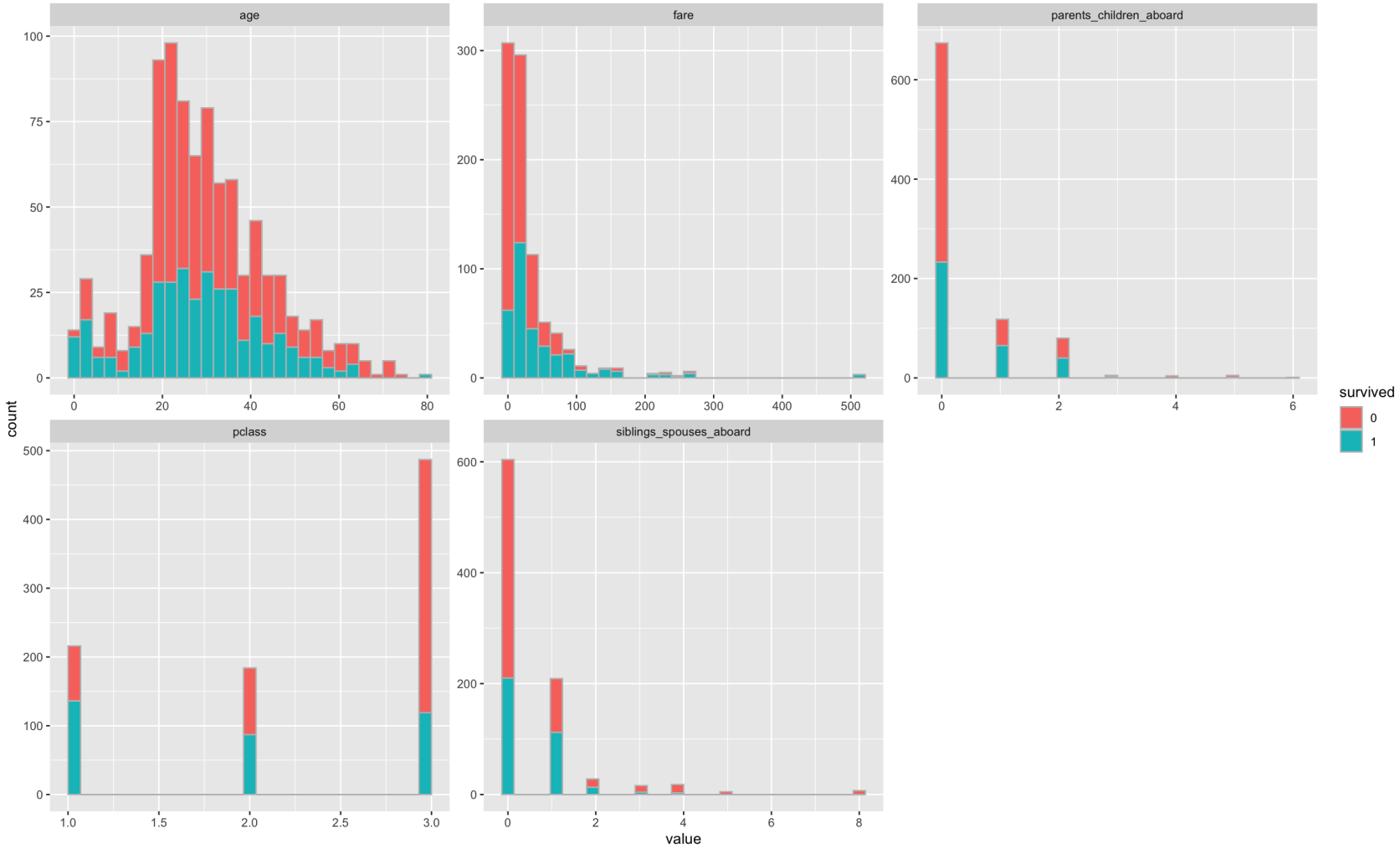
4.2 데이터 시각화

- 탐색적 시각화 예제 : 무엇을 살펴 봐야 할까요?



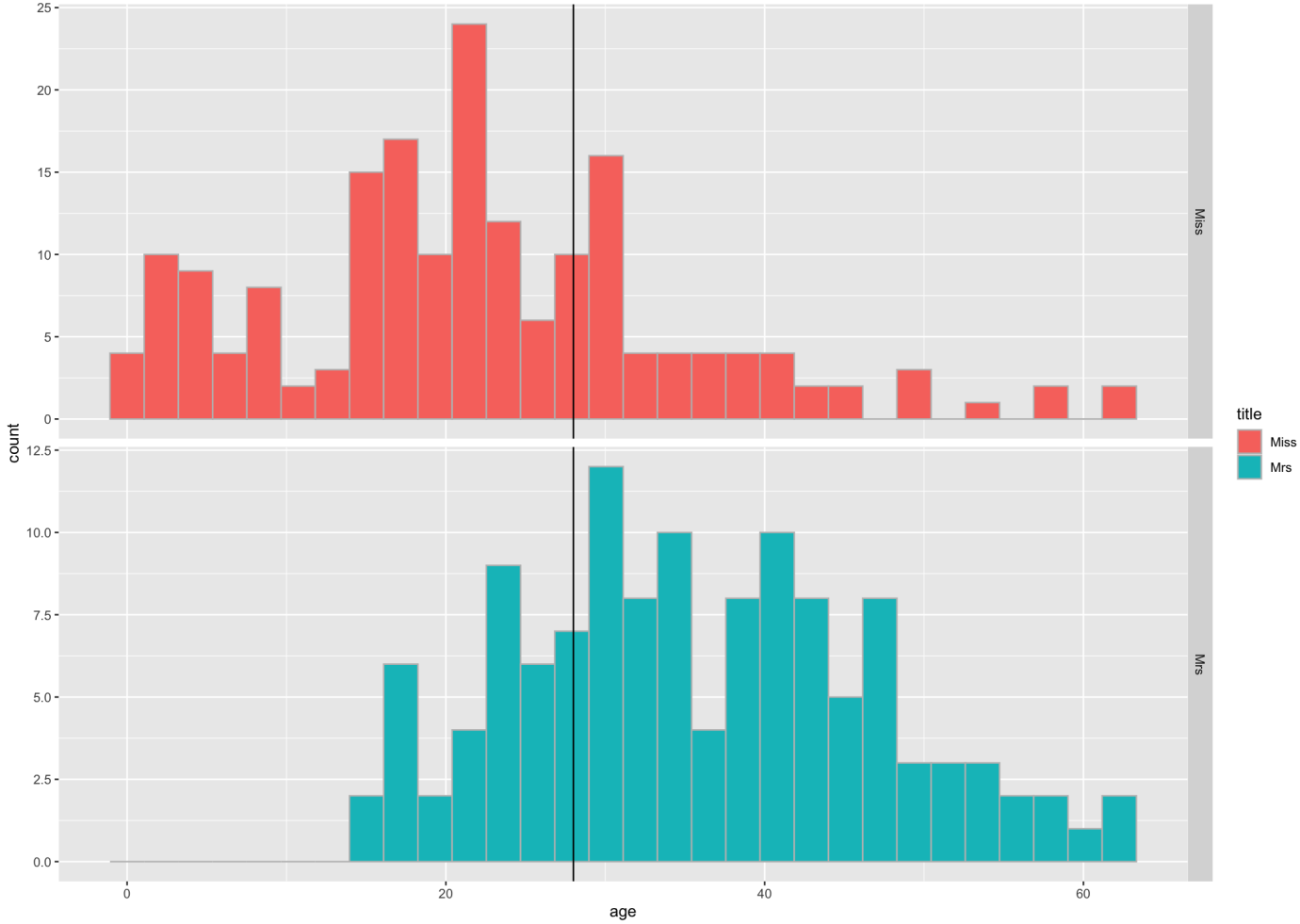
4.3 데이터 시각화

- 탐색적 시각화 예제 : 무엇을 살펴 봐야 할까요?



4.4 데이터 시각화

- 탐색적 시각화 예제 : 무엇을 살펴 봐야 할까요?



5.1 모델링

- 모델링 순서

1. 데이터 분할 : 왜 데이터를 분할하는 것인가요?, 분할 비율의 기준은 무엇인가요?
2. 데이터 전처리 : 전처리 vs. Feature Engineering
3. CV (교차 검증) 정의 : 교차 검증의 이유는? 다른 방법은 없을까요?
4. 모델 정의 : 어떤 모형을 만들 것인가? 몇 개의 모형을 만들 것인가?
5. 모델 튜닝 및 최적 퍼래미터 도출 : 튜닝은 어떻게 할 것인가?
6. 모델 훈련 및 test data 검증 : goodness of fit vs. goodness of prediction
7. 최종 모델 선정

5.2 데이터 분할 및 전처리

- 데이터 분할

  1. 분류 모형에서 데이터 분할을 하기 전에 고려할 사항은 무엇인가요?

  2. 데이터 분할 후 확인할 사항은 무엇인가요?

- 데이터 전처리(Feature Engineering 포함)

  1. 독립변수 간 상관 관계 높은 변수 처리

  2. 왜도 값이 높은 변수 처리

  3. 변수 정규화

  4. 이상 데이터 보정 외

5.3 교차 검증과 모델 정의

- 교차 검증 정의

  1. 교차 검증 방법들 검토

  2. 교차 검증을 위한 fold, 반복의 정도는?

- 모델 정의

  1. 분류 문제에 적용할 모델 고려 (몇 개 모델을 작성할 것인가?)

  2. 모형 간 비교 방법 및 metric 고려

  3. 모델 스펙 정의 (예를 들어 glmnet 모델의 경우 mode, penalty, mixture 정의)
     mode는 분류, penalty의 범위와 L1과 L2의 비율 등

5.4 Hyperparameter Tuning and Model Fit

- hyperparameter 튜닝

  1. grid search or random search?

  2. 튜닝의 폭과 깊이를 어느 정도로 할 것인가?

- 모델 fit, 최종 모델 선정

  1. 개별 모델 내에서 최적 모델 선정

  2. 최적 모델 간 성능 비교

6. 데이터 시각화

- 보고서 시각화
  탐색적 시각화가 모델링 이전 단계인 반면 보고서 시각화는 모델링이 마무리 된 후 의사결정을 위한 시각화로 볼 수 있습니다. 쉽게 말하면 일종의 대시 보드입니다. 주의할 점은 대시 보드를 작성할 때 그 목적에 반드시 부합하는 대시 보드를 작성해야 한다는 것입니다. 기술적 요약이나 데이터 이해를 위한 탐색적 그래프를 대시보드에 올리는 것은 의사결정을 위한 시각화로 볼 수 없습니다. 또한 너무 많은 그래프를 하나의 대시보드에 올리는 일도 집중력을 저하시키는 면이 있으므로 주의해야 할 것입니다.

- 실습 : shiny 패키지

세부 내용은
Decision Insight로 문의 주십시오.

Decision Insight (www.decision-insight.co.kr)
서울 마포구 백범로 31길 21
010 – 5387 – 0300
대표 이후선