HW2_32_A #2 (4조)

1. 데이터 요약

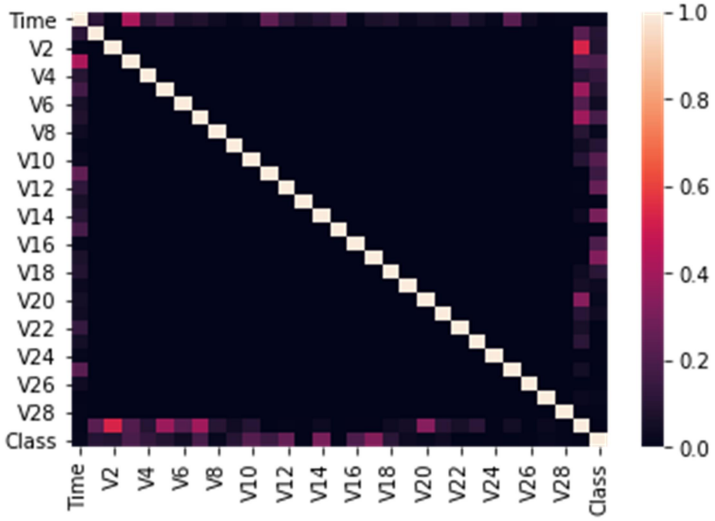
- 데이터 형태: 284,807 row와 31개의 column으로 이루어진 테이블
- Column은 시간 정보를 포함하고 있으며, 28개의 이름을 알 수 없는 column과 amount column, 그리고 class column으로 구성
- 테이블에는 null value를 가지고 있지 않음
- Class가 1인 경우는 0.17% 수준으로 매우 적은 비율을 가지고 있는 불균형 데이터

2. Plot 이미지

(1) 각 컬럼 간의 correlation plot

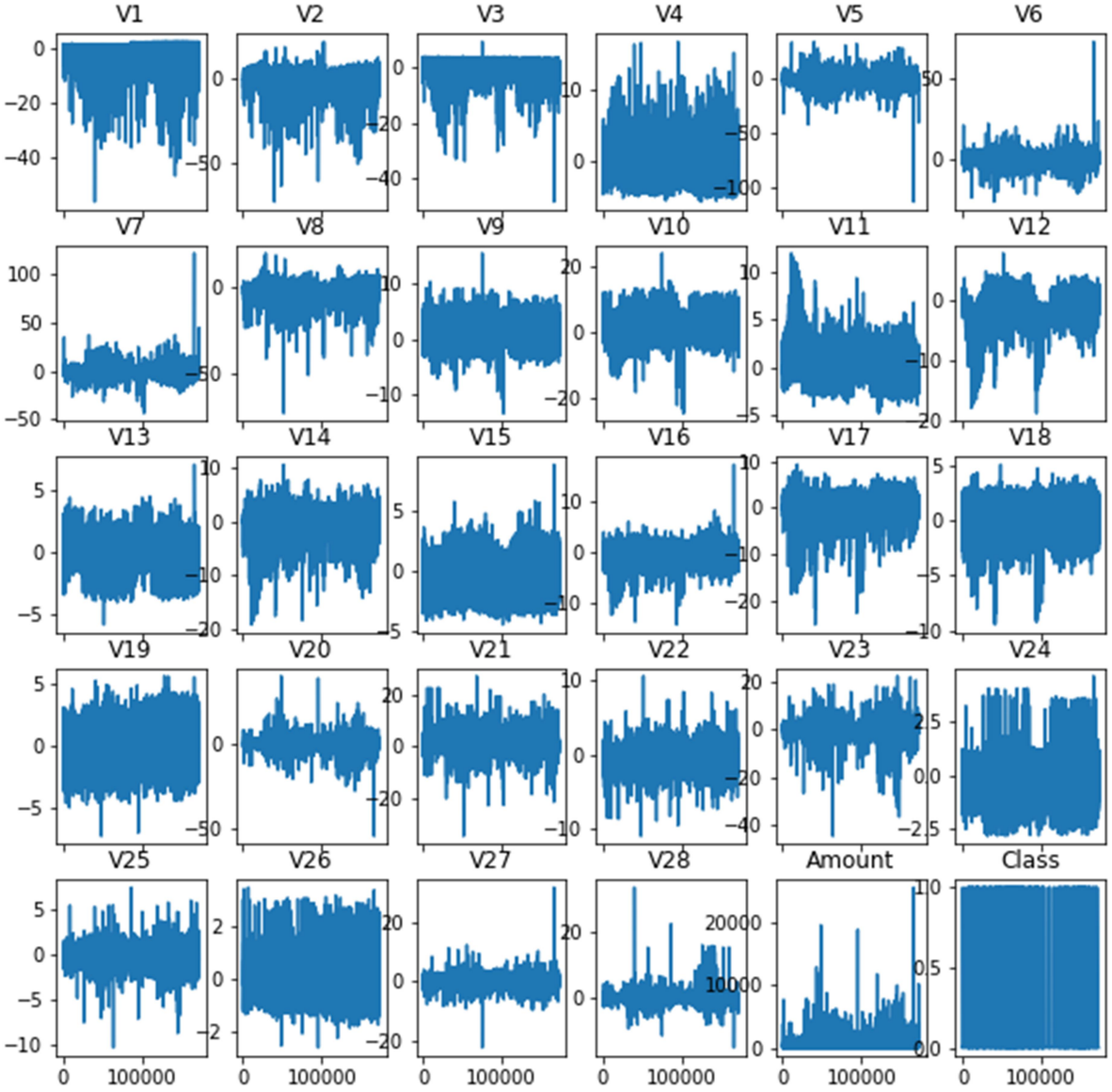
- correlation은 변수들 사이의 상관성을 알 수 있어 좋은 metric으로 사용됨
- column들끼리의 상관성 체크 결과 높은 상관성을 보인 column은 없음

➔ 유의미한 정보 없음



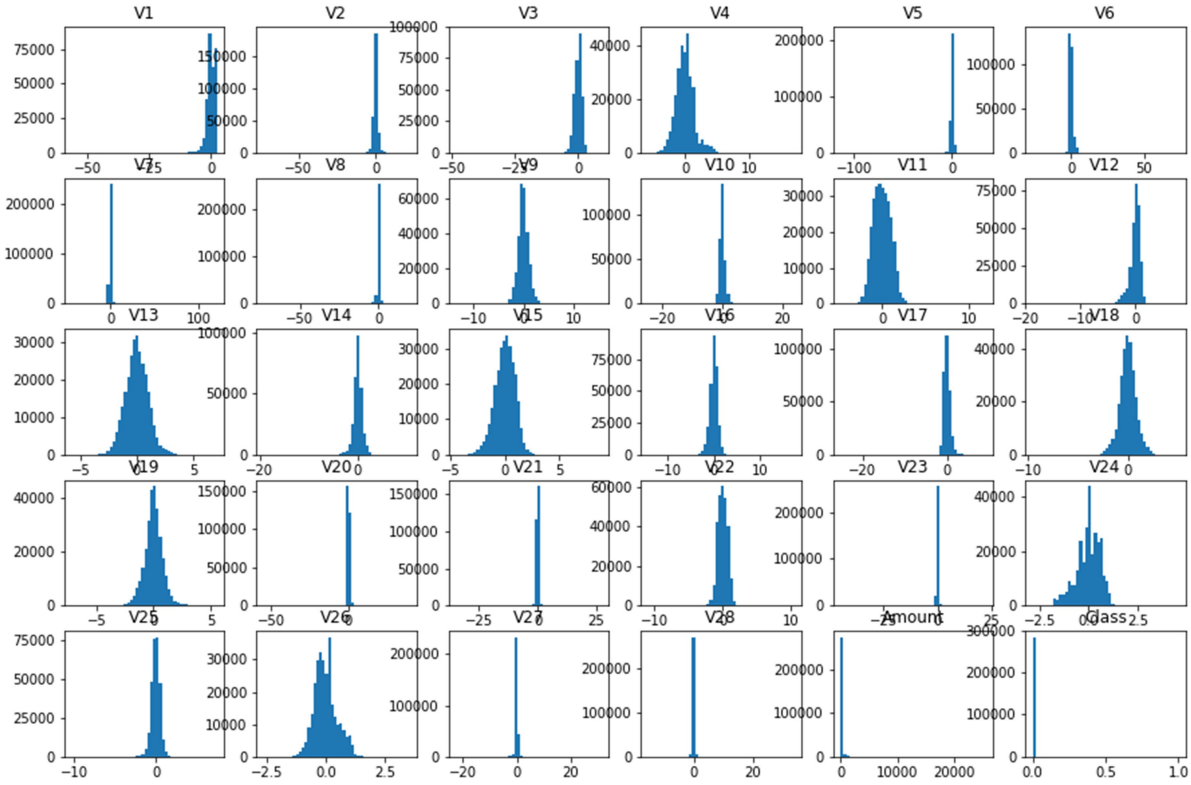
(2) 각 컬럼의 시간별 변화

- 주어진 테이블에 시간에 대한 컬럼이 있으므로 각 컬럼별 시간에 따른 변화 확인
- 시간에 따른 변화 또한 뚜렷한 정보를 제공하지 못함



(3) Histogram

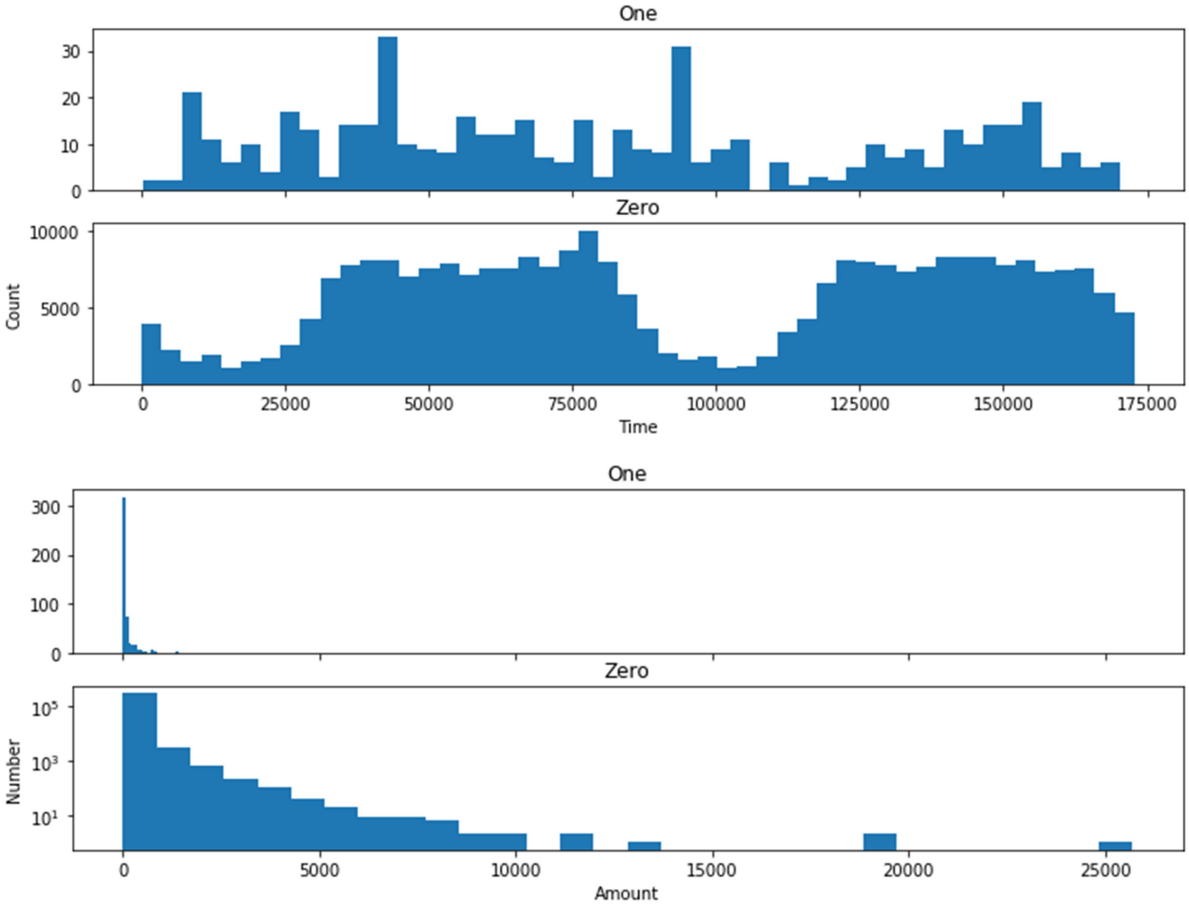
- histogram을 통해 각 컬럼들의 분포를 분석
- 몇몇 컬럼들에서는 정규분포의 형태를 갖는 histogram들이 보임
- Class 구분이 안되었기 때문에 class에 따른 유의미한 정보를 찾지 못함



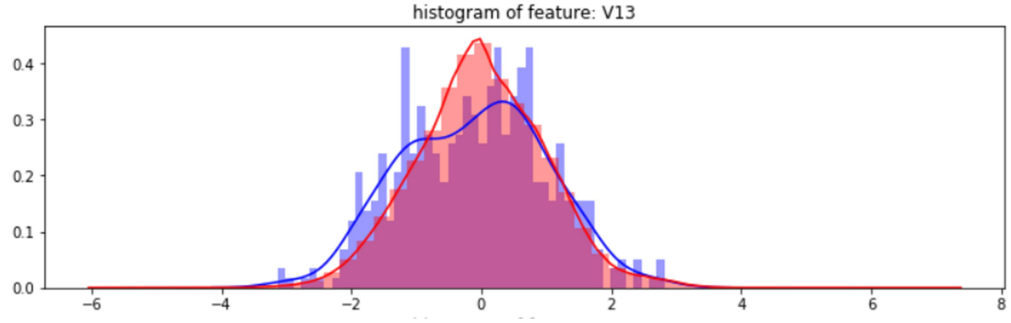
(4) Class에 따른 컬럼별 분포

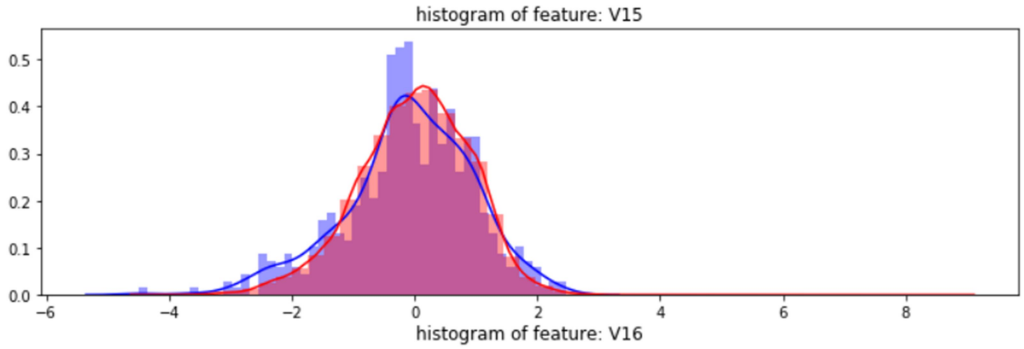
- class가 나타나는 비율이 서로 상이하므로 normalized histogram 분석 필요
- class에 따라 feature (column)의 histogram이 다른 것을 확인할 수 있음
- 즉, feature를 이용하여 class를 구분할 수 있음
- Blue: class 1, Red: class 0

(a) Time과 amount column 분석

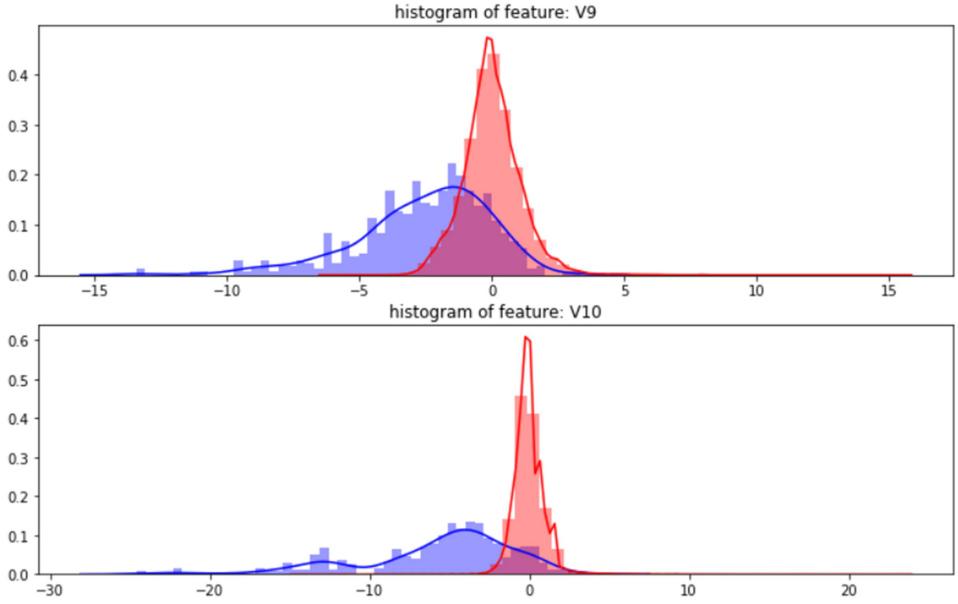


(b) Class에 따라 histogram의 차이가 없는 경우 (일부 케이스만 표기)





(c) Class에 따라 histogram의 차이가 있는 경우 (일부 케이스만 표기)



3. 정리

- 주어진 데이터는 Class 1의 비율이 0.17%인 매우 불균형한 분포를 가지고 있음
- normalize된 histogram을 통해 몇몇 column이 class에 따라 매우 다른 양상을 보이는 것을 확인
- 이런 종류의 column들은 machine learning에 feature로써 활용될 수 있음