Assignment for *Introduction to Data Analysis*

[Important Note]

Assignment #1은 zip 파일에 함께 첨부 된 **wine_a.csv**와 **wine_b.csv** 파일을 사용하여 구현 및 분석을 진행합니다. 결과물은 Jupyter Notebook 코드(.ipynb)와 문제상에서 요구될 경우 Export된 파일을 함께 압축하여 제출하세요.

3. Data Pre-processing

Exercise 3.1

각 csv파일로 부터 데이터를 import하고 ID를 기준으로 하나의 데이터 테이블을 구축한 후 ID 칼럼은 삭제하라.

Exercise 3.2

다음의 조건이 만족하도록 데이터를 정제하고 최종 결과를 *df_final.csv*로 저장하여 코드와 함께 제출하라.

(1) Exercise 3.1에서 만들어진 데이터 테이블에서 시작한다 (ID 삭제 된 상태).

(2) 전체 칼럼 value가 동일할 경우 Duplicate으로 보고 삭제한다.

(3) 각 index의 데이터 칼럼 value 중 하나만 비어 있는 경우 해당 칼럼의 평균값으로 대체하여 사용한다.

(4) 두 개 이상의 value가 비어 있다면 이는 활용할 수 없는 데이터로 간주하고 삭제한다.

(5) 모든 칼럼의 value에 대해서 평균값이 가장 큰 칼럼의 5 standard deviation 보다 큰 value가 있다면 outlier로 간주하고 삭제한다. [Hint: i가 가장 평균값이 큰 칼럼이라면 모든 numerical value는 $\mu_i + 5 * \sigma_i$ 보다 작아야 한다]

(6) 최종 결과 DataFrame에는 ID를 제외한 모든 칼럼이 포함되어 있어야 한다.

(7) 이후 Exercise는 모두 최종 DataFrame을 기준으로 진행한다.

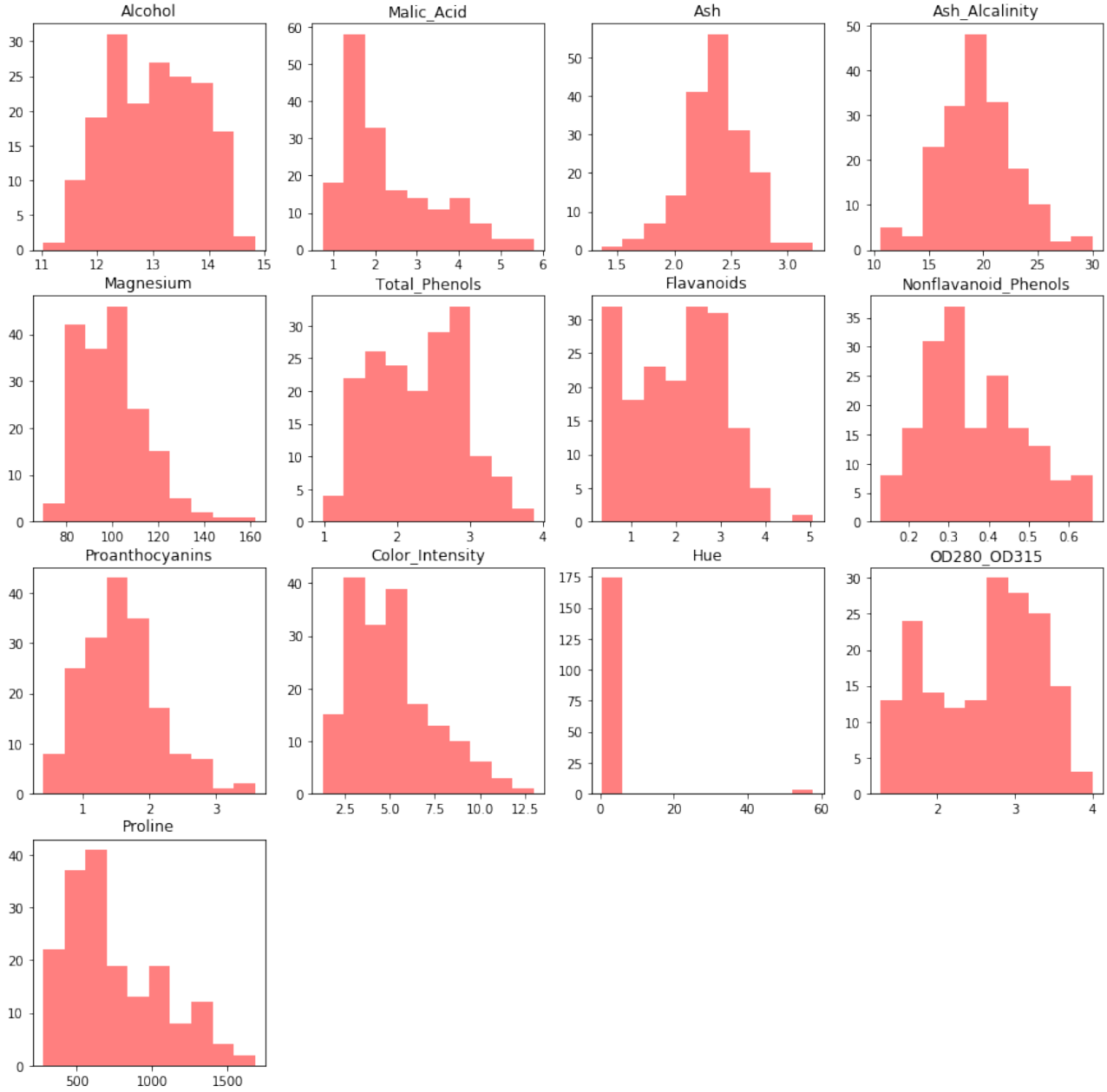
4. Data Visualization

Exercise 4.1

각 칼럼의 분포를 **For-loop**을 활용하여 다음과 같이 Visualize하라.

(1) bin은 지정하지 않는다.

(2) 색은 red(opacity= 0.5)로 지정한다.



Exercise 4.2

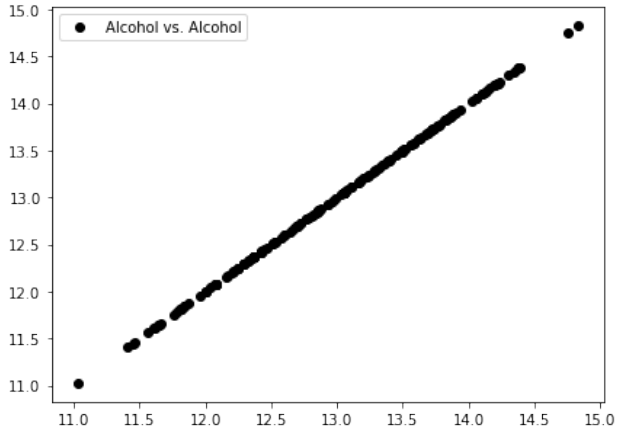
다음의 결과를 Visualize하라.

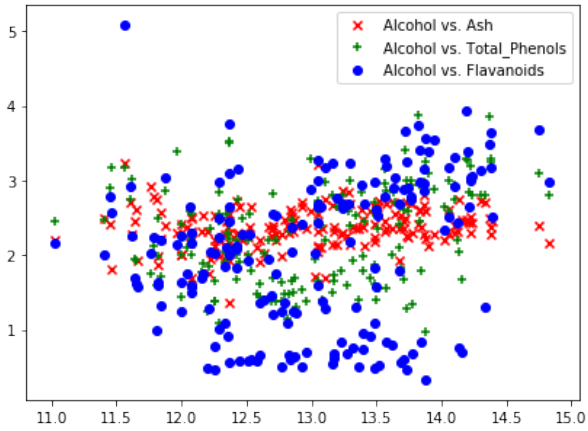
(1) 하나의 캔버스에 2개의 subplot을 그려야 한다.

(2) 좌측은 x축과 y축이 모두 Alcohol이다.

(3) 우측은 x축을 Alcohol로 하고 y축을 여러 다른 칼럼으로 하여 plot한다.

(4) 그림의 크기는 달라도 무방하나 디테일은 정확히 같아야 한다.





5. Dimension Reduction

Exercise 5.1

Type A, B, C의 Descriptive Statistics를 각각 구하라.

Exercise 5.2

전체 데이터(Type 무시)에 대해,

(1) Correlation matrix를 찾은 후 Heatmap을 그려라.

(2) Diagonal(correlation= 1)을 제외하고 가장 강한 positive correlation을 보이는 두 변수를 각기 x, y축으로 하여 Exercise 4.2와 같은 포멧으로 visualize하라.

(3) 결과를 Exercise 4.2의 우측 결과와 비교 분석하라.

Exercise 5.3

전체 데이터(Type 무시)에 대해,

(1) Principal Component 10개를 사용하여 PCA를 실시하라.

(2) 결과를 *df_pca_result.csv*로 저장하여 코드와 함께 제출하라.

(3) PCA 결과에 대한 해석을 Jupyter Notebook의 markdown cell에 서술하라.