データ解析特論　第４回　演習課題

1．重回帰

University of California Irvine では“Machine Learning Repository”という機械学習のベンチマークデータセットを公開している．そのサイトにある“Wine Quality Data Set”（類似名のデータがあるので注意）を用いる．

https://archive.ics.uci.edu/ml/datasets/wine+quality

このデータは，ポルトガルの様々な赤・白のワインの化学成分分析結果による 11 種類の特徴量と主観評価による品質指標が組み合わされたデータである．

“wine-quality”データセットは以下のようにして pandas データフレームとしてロードすることができる：

```
import pandas as pd
try:
    red_table = pd.read_csv("winequality-red.table")
    white_table = pd.read_csv("winequality-white.table")
except:
    url = "https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv"
    red_table = pd.read_table(url, delimiter=";")
    red_table.to_csv("winequality-red.table")
    url = "https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv"
    white_table = pd.read_table(url, delimiter=";")
    white_table.to_csv("winequality-white.table")
```

このデータセットを用いて，以下の課題に取り組みなさい．

(1) 白ワイン，赤ワインそれぞれについて，11 種類すべての特徴量を説明変数とした 1 次式による品質指標の回帰分析を通じて，品質と関連のある特徴量を調査しなさい．結果を元に白ワインと赤ワインではどのような相違があるか考察しなさい．

(2) 特徴量どうしの積（クロスターム）も用いた 2 次式で回帰した場合には，回帰の質は改善するか評価しなさい．

2．分散分析

演習動画でも用いた solder データセットについて,動画では使用していない要因「Solder」,「PadType」,「Panel」について 3 元配置の分散分析を実施し，要因の失敗回数への影響を考察せよ．

レポートには，課題を解くために作成した Python スクリプト，結果，考察を含めること.