

# AIを活かすためのデータ前処理方法の検討

日立製作所  
NECソリューションイノベーション  
日本総合研究所

伊藤 渉太  
大河内 駿太郎  
清野宗一郎

## 開発における問題点

今後AI人材の不足状況はますます深刻化していく見通しであり、AI人材確保には社内育成が第一選択である。しかしながら、現況は環境が整っておらず、中でもAIの前処理は初学者にとってはハードルが高いものである。

## 手法・ツールの適用による解決

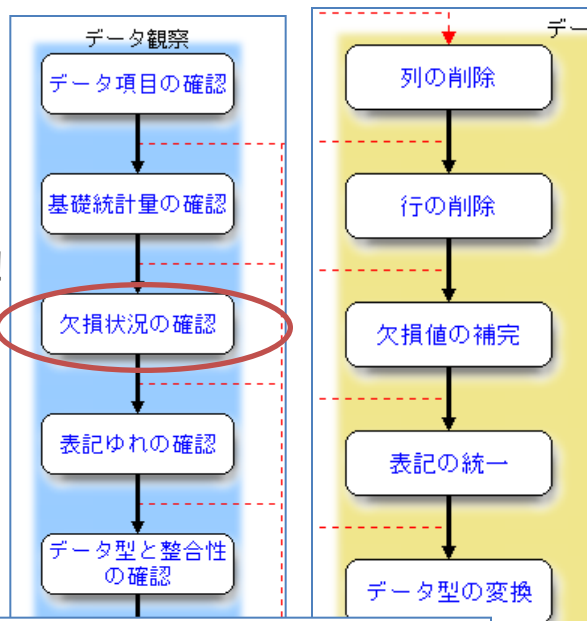
ここでは、データ分析初学者をターゲットとしてペルソナ設定およびカスタマイズによる分析を行うことで、フレームワークを作成し、初学者の生産性を向上し、作業の品質を一定に引き上げることを目指した。

## フレームワークの概要

フレームワークは以下の内容で構成

- ・データ分析の流れ
- ・前処理(データ観察・データ加工)のフローチャート
- ・データ観察作業の詳細(サンプルコード、判断基準)
- ・データ加工作業の詳細(サンプルコード、判断基準)

初学者がどんな作業を実施すべきか  
分かりやすい



### 2.1.3. 欠損状況の確認

データの中に欠損値を含んだ状態では後続に不具合が生じるため、欠損状況を確認する。  
ほとんどのデータが欠損している列は、**列を削除**で列を削除する方がよい。  
また、欠損している行が不要な場合は、**行を削除**で対象の行を削除する。  
欠損値を別の値で補完したい場合は、**欠損値を補完**で欠損値を補完する。

```
#列ごとに欠損値をひとつでも含むか判定。  
#True=欠損値なし、False=欠損値を含む  
print(df.isnull().all())
```

```
##参考：行ごとに欠損値をひとつでも含むか判定  
#print(df.isnull().all(axis=1))
```

```
#列ごとに欠損値の個数をカウント  
print(df.isnull().sum())
```

```
#特定の列に対して、欠損がある行を表示する場合  
df[df["A"].isnull()]
```

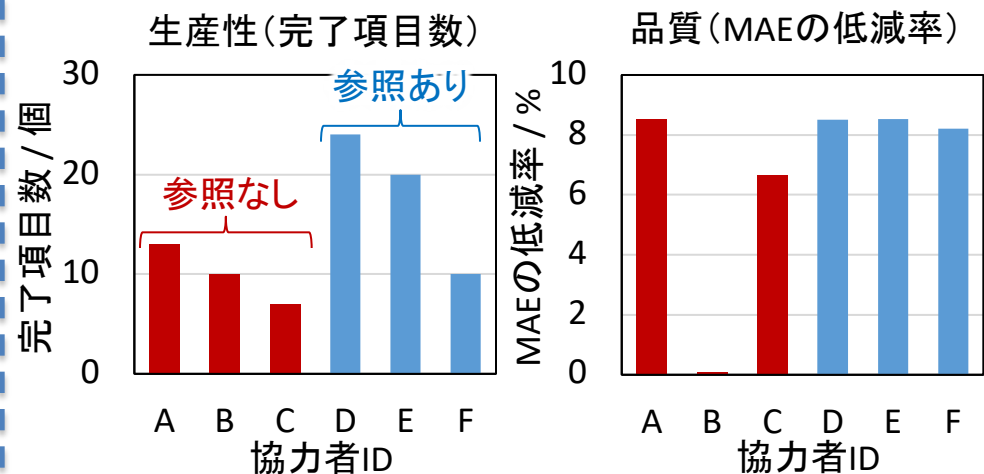
次の作業が  
分かる！

サンプル  
ソースがあるから  
調べる手間が不要

## フレームワークの効果検証(条件)

評価条件	概要
データ	・気象庁の地点気象データセット
予測対象	・翌日の東京の平均気温
評価対象	・データ解析の初学者6名（Python使用経験はあり）
評価方法	・評価対象をフレームワーク参照可否の2グループに分割し、2時間の前処理を依頼
評価指標	・完了項目数(生産性) ・平均絶対誤差(MAE)の低減率(品質)

## フレームワークの効果検証(結果)



## 独立2群のt検定(分散は等しいと仮定)

項目	結果	平均の有意差
生産性	0.0066 (P値) < 0.05 (有意水準)	あり
品質	0.2657 (P値) > 0.05 (有意水準)	ありとは断定できない

フレームワークの参照により生産性が有意に向上