**人員と出荷数の相関を調べ予測するモデルの作成手順**

**目的 : 出荷数と従業員数の相関から適切な人員配置を提案する**

**手順①：データ理解と前処理**

* **データ型の確認:** 必要に応じて変換（例：日付列を datetime型に変換など）
* **文字列をダミー数値化:** (文字列は分析に使えないため)
* **データの分割:** データを訓練・検証データとテストデータに分割 (データ比率8 : 2)
* **欠損値の確認:** （今回は欠損値なし）
* **外れ値の探索と削除:** 散布図を用いて特徴量と出荷数の相関関係や外れ値を確認
* **特徴量の選択:** 相関分析をおこない、出荷数との相関が高い特徴量（従業員数、梱包あり割合、当日出荷割合）を予測モデルの主要な特徴量として選択

**手順②：特徴量エンジニアリング**

* 予測性能の向上を目指して主要な特徴量（従業員数、梱包あり割合、当日出荷割合）の　　2乗項を新たな特徴量として追加。モデルがより複雑な関係性を学習できるようになる

**手順③：モデルの選択と学習（初期：線形回帰）**

* **モデルの選択:** まず、基本的な線形回帰モデルを採用
* **データの標準化:** 特徴量のスケールをそろえ、モデルの学習を安定化させる
* **モデルの学習:** 標準化された訓練データを用いて、線形回帰モデルを学習
* **モデルの評価:** 検証データを用いてモデルの性能を評価

**手順④：モデルの改善検討（ランダムフォレストの導入）**

* **予測の課題:** 線形回帰モデルを用いた予測において、従業員数が増加しても出荷数が　　減少するという、現実的ではない結果が得られる
* **非線形モデルの検討:** 線形回帰モデルでは捉えきれない複雑な関係性を考慮して、強力な非線形モデルであるランダムフォレストの導入を検討
* **ランダムフォレストによる学習:** 特徴量の標準化（特徴量のみ）を行った訓練データを用いて、非線形な関係性や特徴量間の相互作用を捉える能力が高いと期待されるランダムフォレストモデルを学習させ、わずかながら予測結果が改善される

**結果 : モデル作成・分析してわかったこと**　(あくまでも元データは仮だと考慮の上)

* **出荷数と影響度が強いのは、意外にも「当日出荷の割合」だった**
* **従業員の数が増えても、出荷数が増えるわけではない**
* **出荷数に影響を与えている要素が他にもある可能性**