# Iteration: Datenvorbereitung und erste Modellbildung

## Datenverarbeitung:

* **Datenreinigung**: Entfernung von fehlenden Werten aus dem ursprünglichen Datensatz.
* **Duplikatentfernung**: Entfernen doppelter Einträge aus dem Datensatz, um die Qualität des Trainingsdatensatzes zu verbessern.
* **Ausreisserbehandlung**: Entfernen von Extremwerten basierend auf Z-Scores (Schwellenwert 3) für die 'pop' Spalte.

## Modelle:

* **Linear Regression**: Grundlegende lineare Regression ohne Hyperparameter-Anpassung.
* **Polynomiale Regression**: Polynomiales Modell 2. Grades kombiniert mit linearer Regression in einer Pipeline.

## Leistungsmetriken:

* **Linear Regression**: Mittelwert der negativen mittleren quadratischen Fehler über eine 5-fache Kreuzvalidierung.
* **Polynomiale Regression**: Mittelwert des negativen Wurzel-Mittleren-Quadratischen-Fehlers über eine 5-fache Kreuzvalidierung.

# Iteration: Hyperparameter-Optimierung und Modellanpassung

## Datenverarbeitung:

* Keine weiteren Datenverarbeitungsschritte, Weiterführung mit dem bereits vorbereiteten Datensatz.

## Modelle:

* **Polynomiale Regression**: Durchführung von GridSearchCV zur Optimierung der Hyperparameter:
* **Für die polynomiale** **Regression**: Grad der Polynome (2 und 4) und Bias-Einbeziehung.
* **Für den Random Forest**: Maximaltiefe des Baumes (90, 100, 110) und maximale Anzahl von Features (6, 9).

Leistungsmetriken:

* Beste Modelle und Parameter werden ausgegeben, einschliesslich der besten Scores von GridSearchCV für beide Modelle.

# Iteration: Fehleranalyse und Datenvisualisierung

## Datenverarbeitung:

* **Fehlerberechnung:** Berechnung der absoluten Fehler zwischen vorhergesagten und tatsächlichen Werten für das polynomiale Modell.
* **Sortierung und Visualisierung**: Anzeigen der Datensätze mit den grössten Fehlern und Visualisierung der Fehlerverteilung durch Histogramme und Boxplots.

## Modelle:

* Weiterführung mit dem polynomialen Modell, keine neuen Modelle hinzugefügt.

## Leistungsmetriken:

* **Histogramm der Modellresiduen:** Analyse der Residuenverteilung zur Beurteilung der Modellgenauigkeit und Identifizierung möglicher Probleme.

# Iteration: Modellspeicherung

## Datenverarbeitung:

* Keine zusätzlichen Datenverarbeitungsschritte erforderlich.

## Modelle:

* **Speichern des trainierten polynomialen Modells** in eine .pkl-Datei zur späteren Verwendung.