



## California Housing Case Study – Part 2

# Aufgabe 1 : Multiple Regressionsanalyse California Housing

## 1. Datensatz laden und untersuchen:

- Laden Sie das **California Housing Dataset** mithilfe von `fetch_california_housing` aus `sklearn.datasets`.
- Erstellen Sie einen DataFrame und untersuche die ersten Zeilen, um einen Überblick über die Variablen und die Zielvariable `MedHouseVal` (Medianwert der Häuserpreise in Tausend USD) zu erhalten.
- Überlegen Sie welche Variablen potenziell die Häuserpreise beeinflussen könnten. Wähle drei Prädiktoren, die sinnvoll für eine multiple Regressionsanalyse erscheinen.

## 2. Modell erstellen:

- Verwenden Sie die drei theoretisch sinnvollen Prädiktoren um die Zielvariable `MedHouseVal` vorherzusagen.
- Implementieren Sie eine multiple lineare Regression mit `statsmodels.OLS`

## 3. Ergebnisse interpretieren:

- Diskutieren Sie, wie gut das Modell die Varianz der Häuserpreise erklärt...
- ...und welche Variablen einen signifikanten Einfluss auf die Häuserpreise haben
- Offene Diskussion: Welche zusätzlichen Prädiktoren könnten das Modell möglicherweise verbessern?

# Aufgabe 2 : Vorhersage California Housing

## 1. Train-Test-Aufteilung:

- Teilen Sie den Datensatz in Trainings- und Testdaten (80 % Training, 20 % Test).
- Erstellen Sie ein lineares Regressionsmodell mit `scikit-learn.LinearRegression`, und trainiere es mit den Trainingsdaten.

## 2. Vorhersage und Modellbewertung:

- Führen Sie Vorhersagen auf den Testdaten durch
- **Modellevaluation:** Berechnen Sie eine typische Fehlermetrik für Vorhersagen

## 3. Ergebnisse interpretieren:

- Diskutieren Sie, wie gut die Vorhersage ist

# Aufgabe 3 : Qualitative Variablen California Housing

- Fügen Sie dem Datensatz eine **qualitative Variable** auf Basis von AveOccup hinzu: Die kategoriale Variable soll AveOccup in Low, Medium, und High Occupancy separieren.
- Bauen Sie das Modell aus Aufgabe 1 erneut auf, aber fügen Sie die Dummy-Variablen als **neue Variablen** in das Modell hinzu.
- **Vergleichen** Sie das Modell mit und ohne Dummy-Variablen und interpretieren Sie den Einfluss der neuen Variablen auf die Zielvariable MedHouseVal.

# Aufgabe 4: Feature Engineering California Housing

- **Standardisierung und Skalierung:** Standardisieren Sie die Prädiktoren MedInc, AveRooms, und AveOccup mit StandardScaler in scikit-learn.
- Fügen Sie die standardisierten Variablen in das Modell ein und vergleichen Sie die Koeffizienten mit denen der unstandardisierten Variablen.
- **Erstellen neuer Features:** Erstellen Sie eine Variable RoomIncomeRatio, die das Verhältnis von MedInc (medianes Einkommen) zu AveRooms ausdrückt.
- Bauen Sie das Modell erneut auf und fügen Sie RoomIncomeRatio hinzu. Analysieren Sie, wie sich dieses neue Merkmal auf die Vorhersage von MedHouseVal auswirkt.

# Aufgabe 5: Interaktionseffekte California Housing

- **Interaktionsterme berechnen:** Berechnen Sie Interaktionsterme zwischen den Variablen MedInc (medianes Einkommen) und AveRooms (durchschnittliche Raumanzahl pro Haushalt).
- **Modellanalyse:** Interpretieren Sie, wie der Interaktionsterm MedInc\_AveRooms die Zielvariable beeinflusst.

## Aufgabe 6: Polynomial California Housing

- **Erweiterung des Modells mit polynomialen Features:** Fügen Sie einen quadratischen Term für MedInc hinzu, um eine nicht-lineare Beziehung zu modellieren.
- Verwenden Sie PolynomialFeatures in scikit-learn, um das Modell mit  $\text{MedInc}^2$  zu erstellen und anzupassen.
- **Modellvergleich:** Vergleichen Sie die Güte des polynomiellen Modells mit dem einfachen linearen Modell.
- Diskutieren Sie das Risiko von Overfitting und vergleichen Sie die Güte auf den Testdaten mit der des linearen Modells.

## Aufgabe 7: Stepwise Regression California Housing

- **Schrittweise Prädiktorauswahl:** Führen Sie eine Vorwärts- und Rückwärtsselektion für das Modell durch, um die wichtigsten Prädiktoren zu identifizieren.
- Nutzen Sie statsmodels für die Schrittweise Regression und dokumentieren Sie die Reihenfolge, in der die Prädiktoren hinzugefügt oder entfernt werden.
- **Modellvergleich:** Vergleichen Sie das Modell mit allen Prädiktoren und das reduzierte Modell, das aus der Feature Selection hervorgeht.
- Analysieren Sie die Auswirkungen auf die Modellgüte und die Interpretierbarkeit.

## Aufgabe 8: QQ-Plot California Housing

- **Erstellung eines Q-Q-Plots:** Verwenden Sie `scipy.stats.probplot` oder `statsmodels` für einen Q-Q-Plot der Residuen.
- **Interpretation der Normalverteilung:** Analysieren Sie die Normalverteilung der Residuen und diskutieren Sie, wie Abweichungen die Modellgüte und Interpretation beeinflussen.
- Diskutieren Sie, ob die Residuen im Modell normalverteilt sind und welche Transformationsmöglichkeiten bestehen, falls Abweichungen festgestellt werden

## Aufgabe 9: Variance Inflation Factors California Housing

- **Multikollinearität analysieren:** Berechnen Sie den Variance Inflation Factor (VIF) für jedes Feature.
- Bestimmen Sie die Prädiktoren mit hohen VIF-Werten und diskutieren, welche Variablen entfernt oder kombiniert werden sollten, um Multikollinearität zu reduzieren.