

**Duale Hochschule Baden-Württemberg Mannheim**

**Data Exploration Projekt:**

**“Inside Explorers”**

House Price Prediction

Gruppenmitglieder: Felix Hüsgen 3583266

Lennart Fertig 8602307

Pascal Schmidt 8133405

Studiengang: Wirtschaftsinformatik – Data Science

Kurs: WWI19DSB

Dozent: Sebastian Schön

Simon Poll

Inhaltsverzeichnis

[1. Thema und Motivation 3](#_Toc73867078)

[2. Auswahl des Datensatzes 3](#_Toc73867079)

[3. Herangehensweise 4](#_Toc73867080)

Inhaltsverzeichnis:

- Thema und Motivation

- Related Work (welche wissenschaftlichen Publikationen gibt es zu diesem Thema bereits? )

- Verwendete Technologien und Bibliotheken (z. B. scikit learn , tensorflow,…)

- Präsentation der Ergebnisse

- Kritische Bewertung der Ergebnisse („ lessons learned “: Was hat (nicht) funktioniert

- Anmerkungen zum Quellcode im Anhang (wie ist der Code auszuführen und was gibt es zu beachten?)

# Thema und Motivation

Seit einigen Jahren steigen bekanntlich die Immobilienpreise in den USA wie auch in Europa rasant an. Vor allem große Städte sind von den Preissteigerungen betroffen. Generell sind Hauspreise von hohem Interesse für verschiedenste Shareholder und somit gerade die Voraussage dieser Preise von hoher Bedeutung.

Ziel dieses Data Exploration Projektes ist es , Hauspreise auf Basis verschiedener Eigenschaften durch den Einsatz von Machine Learning Modellen vorauszusagen. Da keine Preisklasse, sondern ein genauer Preis prognostiziert werden soll, wird im Projekt auf Regressionsmodelle zurückgegriffen. Durch solche Vorhersagen können Shareholder eine bessere Informationslage erlangen und somit bessere Kaufentscheidungen treffen. Aspekte wie der optimale Kaufzeitpunkt oder auch Renovierungsvorhaben können so besser geplant werden.

Bei der Vorhersage sollen typische Aspekte der Bewertung einer Immobilie wie Einrichtung und Lage berücksichtigt werden. Beispielsweise kann die Größe der Wohnfläche oder auch die Anzahl der Badezimmer einen großen Einfluss auf den Hauspreis haben.

# Auswahl des Datensatzes

Bei der Auswahl eines geeigneten Datensatzes für das beschriebene Ziel hat sich die Gruppe für einen Datensatz von der Online-Community Kaggle entschieden[[1]](#footnote-1). Der Datensatz beinhaltet Hausdaten von 21600 Häusern inklusive Verkaufspreisen in King County bzw. Seattle in den USA. Es sind 20 verschiedene Attribute wie Wohnfläche, Baujahr oder Anzahl der Badezimmer für jedes Haus gegeben.

Trotz der eher kleinen Anzahl an Daten ist der Datensatz aufgrund seiner Vollständigkeit und der vielen Attribute sehr gut für das Projekt geeignet. Dadurch, dass die Häuser einer einzelnen großen Stadt betrachtet werden, erhofft sich die Gruppe bessere Ergebnisse als beispielsweise bei der Betrachtung von Häusern eines ganzen Landes oder Kontinents. In einer einzelnen Stadt sind bestimmte Korrelationen möglicherweise stärker vorhanden.

# Herangehensweise

Die Gruppe hat sich dazu entschieden, den Python-Code in Google Colaboratory zu entwickeln, da hier gemeinsam an Notebooks gearbeitet werden kann und einige benötigte Module wie scikit-learn schon vorinstalliert sind. Somit kann der Code darüber hinaus schnell und unkompliziert mit anderen Personen geteilt werden.

Zunächst wird der Datensatz als csv-Datei über Google Drive in die Colaboratory-Umgebung geladen und in ein pandas-Dataframe verpackt. Das anschließende Vorgehen beginnt mit einer Datenanalyse, bei der die Daten vorrangig auf Zusammenhänge untersucht werden. Danach werden die Daten für den Machine Learning Einsatz vorbereitet. Es sollen verschiedene Modelle zur Regression trainiert werden, die dann anschließend evaluiert und gegebenenfalls optimiert werden.

1. Der Datensatz kann unter folgendem Link aufgerufen werden: <https://www.kaggle.com/harlfoxem/housesalesprediction> (Aufruf: 05.06.2021) [↑](#footnote-ref-1)