****

**Duale Hochschule Baden-Württemberg Mannheim**

**Data Exploration Projekt:**

**‘Inside Explorers‘**

House Price Prediction

Studiengang Wirtschaftsinformatik

Data Science

**Gruppenmitglieder:** Felix Hüsgen 3583266

Lennart Fertig 8602307

Pascal Schmidt 8133405

**Studiengang**: WWI19DSB

**Studiengangsleiter**: Prof. Dr. Bernhard Drabant

**Dozenten:** Sebastian Schön,

Simon Poll

**Kurs:** Data Exploration

**Bearbeitungszeitraum**: 09.06. – 06.07.21

Inhaltsverzeichnis

[1. Thema und Motivation 3](#_Toc76461964)

[2. Grundlagen 4](#_Toc76461965)

[2.1. Auswahl des Datensatzes 4](#_Toc76461966)

[2.2. Herangehensweise und verwendete Technologien 4](#_Toc76461967)

[2.3. Related Work 5](#_Toc76461968)

[3. Ergebnisse 6](#_Toc76461969)

[3.1. Datenanalyse, Visualisierung und Vorverarbeitung 6](#_Toc76461970)

[3.2. Präsentation der Ergebnisse 6](#_Toc76461971)

[3.3. Kritische Bewertung 6](#_Toc76461972)

[Anhang (Anmerkungen zum Quellcode) III](#_Toc76461973)

# Thema und Motivation

Seit einigen Jahren steigen 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.

# Grundlagen

## 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. 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 und verwendete Technologien

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.

## Related Work

Während der Bearbeitung des Projektes wird unteranderem Bezug auf das Buch „Praxiseinstieg Machine Learning mit Scikit-Learn und TensorFlow: Konzepte, Tools und Techniken für intelligente Systeme“ von Aurélien Géron genommen[[1]](#footnote-1). Dies eignet sich besonders gut für die Bearbeitung des Data Exploration Projekts, da es sich direkt an angehende Data Scientists richtet und den Einsatz von einfachen und effizienten Werkzeugen zum Implementieren eines Machine-Learning-Projekt vermittelt.

Besonders wird sich dabei auf Kapitel 2 „Ein Machine-Learning-Projekt von A bis Z“ bezogen, welches der Gruppe anhand eines Beispiels die wichtigsten Schritte in einem Data Science Projekt vermittelte. Dabei beschreibt Aurélien Géron, angefangen von der Beschaffung der Daten über ihre Erkundung und Visualisierung, sowie der Vorbereitung der Daten für den Machine-Learning Algorithmus, die einzelnen Etappen für den erfolgreichen Ablauf eines Machine-Learning-Projektes.

# Ergebnisse

## Datenanalyse durch Visualisierung und Vorverarbeitung

Nachdem der Datensatz erfolgreich über Google Drive in die Colaboratory-Umgebung geladen und in ein pandas-Dataframe verpackt wurde, wird zunächst die Datenstruktur analysiert. Dazu nutzt die Gruppe, neben der Methode *info()*, die eine schnelle Beschreibung des Datensatzes liefert, die Methode *describe()*, um die numerischen Merkmale zusammenfassen zu können. Des Weiteren wird durch das Erstellen von Histogrammen, tiefere Erkenntnisse über die Mengenverteilung jedes nummerischen Attributes, bezogen auf die einzelnen Immobilien, gewonnen (siehe Abbildung 1). Im nächsten Schritt wird mithilfe der Methode *corr()* der Korrelationskoeffizient berechnet, um die Korrelation jedes Merkmals mit dem Preis der Immobilien zu erhalten. Dieser wird neben einer Heatmap (siehe Abbildung 2), auch mithilfe weiterer Scatter Plots visualisiert. Durch das zur Verfügung stehen von geografischen Informationen (Breite und Länge), wird weiter die geografische Lage einzelner Immobilien dargestellt. Jede Immobilie wird durch einen Datenpunkt beschrieben, wobei die Farbe eines jeden Datenpunktes den Preis der Immobilie verdeutlicht. Dafür wird die vordefinierte Farbskala (seismic) verwendet, wobei niedrigere Hauspreise mithilfe der Farbe Blau, höhere Hauspreise durch die Farbe Rot, visualisiert werden (siehe Abbildung 3).

Nachdem die Gruppe den Datensatz durch Analysen und Visualisierungen erkundet hat, stehen weite Erkenntnisse zur Verfügung, um einen Trainings- und Testdatensatz zu generieren. Dabei werden zunächst alle Immobilien, die einen Preis von über 4 Millionen Dollar haben aus dem Datensatz entfernt, um einen zu großen Einfluss auf die Model Performance zu verhindern. Um einen Testdatensatz zu erstellen, wählt die Gruppe 20 Prozent des reduzierten Datensatzes aus, um einen „Train-Test“ Split durchzuführen.

Im nächsten Schritt, werden die Daten für den Machine-Learning-Algorithmus vorbereitet. Dabei sind durch den gegebenen Datensatz alle Datenpunkte je Merkmal gegeben, wodurch ein setztes des Datenpunktes auf einen bestimmten Wert oder ein gar löschen des Merkmals aus dem Datensatz nicht notwendig ist. Da zusätzlich in dem Datensatz keine kategorischen Merkmale gegeben sind, ist das Konvertieren von Einträgen zu nummerischen Einträgen nicht erforderlich. Dennoch wird der Datensatz durch das herausnehmen von den Merkmalen ‚date‘, ‚zipcode‘, ‚id‘ und ‚price‘ (da ‚price‘ das Label ist) weiter angepasst, bevor dieser Skaliert wird. Eine Skalierung des Datensatzes ist erforderlich, damit der Machine-Learning-Algorithmus besser aus diesem umgehen und aus diesem lernen kann. Dazu wird eine Pipeline initialisiert, um die Daten mithilfe des ‚StandartScalers‘ aus der Scikit-Learn Bibliothek, skalieren zu können. Dieser organisiert die Abfolge der nötigen Transformationsschritte, um einen skalierten Datensatz zu erhalten.

Am Ende dieses Schrittes steht der Gruppe ein vorbereiteter Trainingsdatensatz ohne ‚Labeling‘ zur Verfügung.

## Präsentation der Ergebnisse

## Kritische Bewertung

# Anhang (Anmerkungen zum Quellcode)

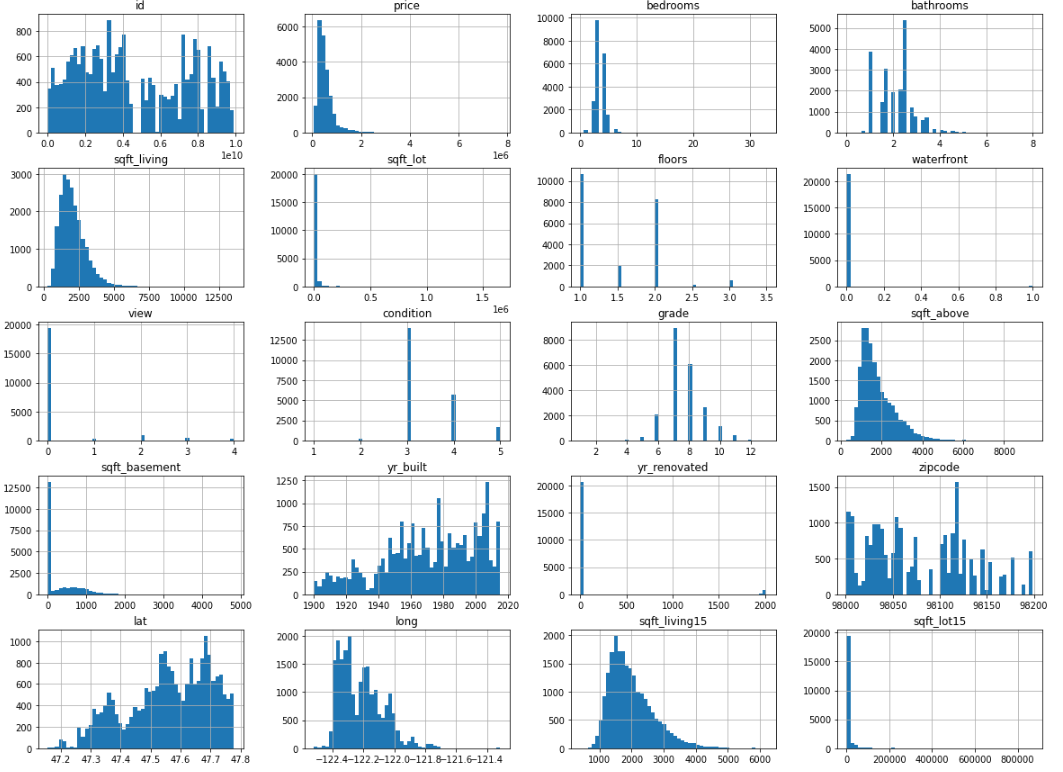


Abbildung 1: Histogramme zur Visualisierung der Mengenverteilung der einzelnen Merkmale

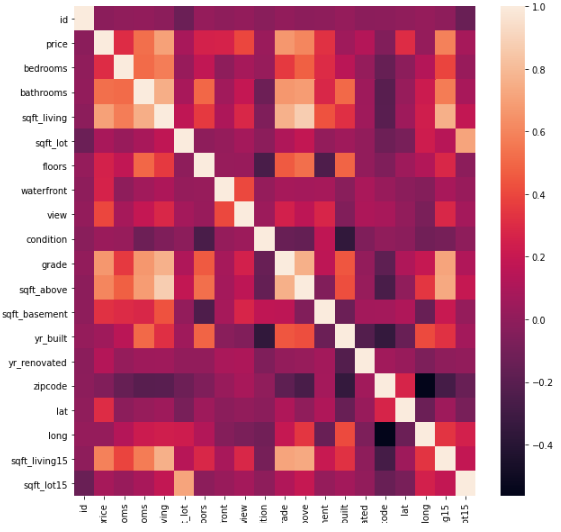


Abbildung 2: Heatmap zur Visualisierung der Korrelation der einzelnen Merkmale.

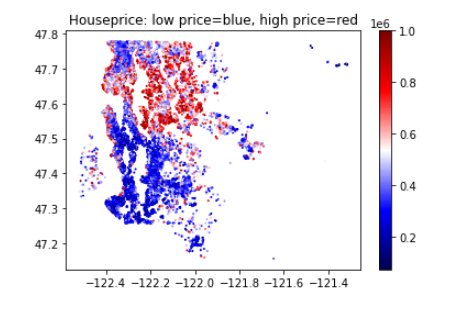


Abbildung 3: Visualisierung der geografischen Lage und des Preises einzelner Immobilien in Seattle.

1. Zum Buch: <https://oreilly.de/produkt/praxiseinstieg-machine-learning-mit-scikit-learn-keras-und-tensorflow/> (Zugriff am 27.06.21) [↑](#footnote-ref-1)