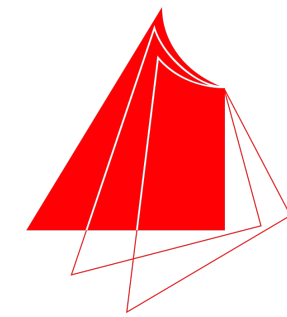


KI Summer School

Übung 1: Pandas



Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

Prof. Dr. Patrick Baier

Lernziel

In der ersten praktischen Übung geht es darum, mit Hilfe von Python, Jupyter und Pandas einen Datensatz zu laden und zu analysieren.

Aufgabe 1

1. Starten Sie ein neues Jupyter Notebook und laden Sie den `automobile.csv` Datensatz aus dem `data` Ordner in einen Pandas DataFrame.
2. Schauen Sie sich die Daten an und versuchen Sie zu verstehen was die Daten abbilden.
4. Geben Sie folgende Daten aus:
 1. Eine Liste aller Hersteller.
 2. Wie oft welcher „body-style“ im Datensatz vorhanden ist.
 3. Die fünf Autos mit dem niedrigsten Preis. ([Hint](#))
 4. Die fünf Autos mit dem höchsten Preis.

Aufgabe 2

1. Was ist der durchschnittliche Preis mit Bezug auf den „body-style“ eines Autos.
2. Welches sind die drei „body-styles“ mit den durchschnittlich höchsten Preisen?
3. Welche Companies stellen „hardtop“ als „body-style“ her?
4. Geben Sie alle Autos aus deren Herstellername mit „a“ oder „b“ beginnt ([Hint](#)).
5. Erstellen Sie je einen Scatterplot mit:
 1. `x="horsepower"` und `y="average-mileage"`
 2. `x="horsepower"` und `y="price"`

Welche Zusammenhänge lassen sich erkennen?

Aufgabe 3

1. Zusätzlich stehen nachfolgende Informationen zur Verfügung. Erstellen Sie einen zweiten DataFrame, der diese Informationen enthält ([Hint](#)):

company	country
alfa-romero	ITA
audi	GER
bmw	GER
volvo	SWE
nissan	JAP
mazda	JAP

2. Verbinden Sie die zwei DataFrames über einen *merge* ([Hint](#)). Stellen Sie sicher, dass die Datenpunkte welche keine Länderangabe haben nicht verloren gehen.
3. Geben Sie in einem DataFrame aus, wieviel Fahrzeuge pro Land vorhanden sind und wie hoch der Durchschnittspreis pro Land liegt.

Bonus Aufgaben

Folgende Aufgaben sind als Bonus gedacht (für diejenigen die mit den vorherigen Aufgaben schon durch sind). Um Sie zu lösen muss selbstständig nach den benötigten Methoden im Internet gesucht werden.

1. Entfernen Sie die Spalte „body-styles“ vom DataFrame.
2. Erstellen Sie einen neuen DataFrame der 10% zufällig ausgewählte Zeilen aus dem ursprünglichen DataFrame enthält.
3. Fügen Sie die Zeilen des neuen DataFrames and den ursprünglichen DataFrame unten ran.
4. Zählen Sie die Anzahl doppelter Einträge in diesem DataFrame.

Fragen?