

# Projektarbeit im 1. Sem. Informatik

Studiengang Wirtschaftsingenieurwesen, DHBW Stuttgart (Prof. Heuser)

## Projekt 4 („Fahrzeugdaten Daimler Truck“) – mittelschwer-schwer

Als Grundlage zu dieser Projektarbeit dienen folgende Dateien, welche uns freundlicherweise vom Data Scientist Jannik Wiessler, Fa. Daimler Truck AG Stuttgart, zur Verfügung gestellt wurden. Es handelt sich bei den Daten um fiktive, aber realitätsnahe Unternehmensdaten sowie um eine echte Einstiegsaufgabe für zukünftige Data Science Mitarbeiter/innen von Daimler Truck. Zunächst sollten Sie die ZIP-Datei `vehicle_data_project_exercise.zip`, welche Sie in *Moodle > Vorlesungsmaterialien > Vorlesung\_Heuser > Projektarbeit* vorfinden, herunterladen und in Ihrem lokalen Verzeichnis entpacken. Danach werden Sie dort folgendes vorfinden:

- `vehicle_data_project_exercise\main_ex.py` (zu ergänzende Main-Routine)
- `vehicle_data_project_exercise\import_data_ex.py` (zu ergänzende Datenimport-Routine)
- `vehicle_data_project_exercise\pipeline_ex.py` (zu ergänzende Pipeline-Routine)
- `vehicle_data_project_exercise\analysis_ex.py` (zu ergänzende Analysis-Routine)
- `vehicle_data_project_exercise\data_specs.json` (zu ergänzende Data Specification)
- `vehicle_data.xlsx` (Quell- bzw. Rohdaten)
- `requirements.txt` (Requirements)
- `praxisprojekt.pdf` (Zusatzerläuterungen zum Projekt 4)

Ein kurzes Erklärvideo, in dem die Rohdaten besprochen werden, stellen wir Ihnen zusätzlich bereit:

<https://www.youtube.com/watch?v=0dNxU1vBV5M&list=PLEEzqFSANCS0v3C-cfYXVslcfQ5ldICV>

Die Daten, welche in dieser Challenge verwendet werden sollen, finden Sie in der Datei `vehicle_data.xlsx`. Bitte beachten Sie, dass die Daten zunächst aufbereitet werden müssen, also noch nicht in einer Qualität vorliegen, die eine finale Insight-Generierung ermöglicht.

Die Quelldatei `vehicle_data.xlsx` beinhaltet Daten zu 500 Fahrzeugen:

Im Reiter `sales_code` finden Sie 500 Trucks deren eindeutige Fahrzeugidentifikationsnummer (FIN) als hash in der Spalte `h_vehicle_hash` dargestellt ist. Jedes Fahrzeug ist aufgeschlüsselt nach Produktionsdatum, Land, in welches das Fahrzeug verkauft wurde und `sales_code_array`. Letztere Spalte beinhaltet Codes, welche die genaue Zusammensetzung des Fahrzeuges beschreiben. Im Reiter `engines` finden Sie 9 `sales` Codes, die Aufschluss darüber geben, welcher Motor im entsprechenden Fahrzeug verbaut ist.

Mit der Tabelle `vehicle_hash` können Sie die gehashte FIN zurück in die originale Darstellung mappen.

In der `requirements.txt` finden Sie die benötigten Packages: (`pip install -r /path/to/requirements.txt`)

Die Challenge teilt sich in zwei Aufgabenteile ein. Der erste Teil zielt auf ihre Softwareentwicklungsfertigkeiten ab. Halten Sie bei der Code Gestaltung die gängigsten Clean Code Regeln ein und verfolgen Sie nach Möglichkeit eine testgetriebene Entwicklung.

Im zweiten Aufgabenteil sollen Sie verschiedene Analysen durchführen. Verwenden Sie dazu die Daten, welche Sie in Aufgabenteil eins aufbereitet haben. Bevor Sie starten, bietet es sich an sich zunächst mit den Datenätzen vertraut zu machen. Verschaffen Sie sich einen Überblick und erkunden Sie die Zusammenhänge der Datensätze.

## Ihre Aufgabe im Team

Halten Sie sich bei der Bearbeitung der Aufgabe zunächst immer an die allgemeinen Projektvorgaben aus der Datei „Projektarbeit“!

Spezielle Projektaufgaben:

### Aufgabe 1: Data Engineering

Schreiben Sie eine ETL Pipeline zur Datenaufbereitung. Gehen Sie dabei wie folgt vor.

- Daten laden
- Daten bereinigen und aufbereiten
- Daten zusammenführen
- Gesamttabelle bestehend aus folgenden Spalten abspeichern:
  - o `fin`
  - o `production_date`
  - o `country`
  - o `sales_code_array`

### Aufgabe 2: Data Science

Analysieren Sie den die Daten, indem Sie folgende Fragestellungen auswerten. Visualisieren Sie ihre Ergebnisse.

- Welches sind die top drei Länder, in die wir zwischen 01.01.2014 und 31.12.2020 am meisten Fahrzeuge verkauft haben.
- In welchem dieser Jahre haben wir insgesamt am meisten Fahrzeuge verkauft?
- Welche FIN hat das zeitlich erste verkaufte Fahrzeug.
- Wie viele Fahrzeuge wurden zwischen 01.01.2017 und 01.01.2021 mit OM934, OM936, OM470 und OM471 Motoren verkauft.
- Welche Fahrzeuge (FIN) wurden zwischen 01.01.2017 und 01.01.2021 und mit OM936 Motor nach Neuseeland verkauft.

## Hilfe

Für sämtliche graphische Funktionen der Libraries Matplotlib, Pandas und darüber hinaus stehen Ihnen viele nützliche Websites als Hilfe zur Verfügung. Sollten Sie also Funktionsdetails der Bibliotheken oder Beispiel-Anwendungen mit Hilfe der Bibliotheksfunktionen nachschlagen wollen, so sei Ihnen folgende URL empfohlen:

- <https://docs.python.org/3.9/> (sämtliche in Python integrierte Funktionen inkl. Library Reference und Tutorial; Python Version 3.9 oder neuere)
- <https://matplotlib.org/stable/tutorials/introductory/pyplot.html> (Pyplot)
- [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/10min.html](https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html) (Pandas)
- <https://realpython.com/beginners-guide-python-turtle/> (Turtle)
- ...

Häufig ist es auch ratsam, mit Hilfe von Google nach Hilfen im Web zu suchen.