Problem Set 2 – Erste Schritte mit ggplot

Vorbereitung

- ☐ Laden Sie den Datensatz **ToyotaCorolla.csv** von GitHub herunter (das ist der gleiche Datensatz wie in Problem Set 1) und speichern Sie den Datensatz in Ihrem Arbeitsverzeichnis.
- ☐ Erstellen Sie ein neues R-Skript mit dem Namen **Problem_Set_2.R**
- ☐ Ergänzen Sie in ihrem R-Skript folgende erste Zeile: **rm(list=ls())**
- ☐ Lesen Sie den Datensatz **ToyotaCorolla.csv** über das Menü von RStudio ein und vergeben Sie für diesen DataFrame den Namen **cars_df**. Vergessen Sie nicht, den generierten Code in Ihr R-Skript zu kopieren.
- ☐ Betrachten Sie den Datensatz mit dem Befehl **View(...)**.
- ☐ Laden Sie das Paket tidyverse via des Befehls **library(Paketname)**.
- ☐ Prüfen Sie, ob der Datensatz Beobachtungen mit fehlenden Informationen hat. Korrigieren Sie dies bei Bedarf.
- ☐ Nutzen Sie die Funktion **dim( )**, um die Anzahl an Beobachtungen und Variablen des Datensatzes zu bestimmen.
- ☐ Speichern Sie Ihr R-Skript und führen Sie Ihr gesamtes Skript noch einmal aus und prüfen Sie, ob Ihr Vorgehen replizierbar ist.

Analyse

Sie interessieren sich für den Zusammenhang zwischen dem Preis eines Toyotas (**Price**) und der Anzahl der gefahrenen Kilometer (**KM**).

- ☐ Berechnen Sie den Mittelwert für beide Variablen und speichern Sie die beiden Mittelwerte in zwei R-Objekten mit dem Namen **Variablenname_mw** ab.
- ☐ Nutzen Sie die Funktion **ggplot( )** um ein Plot Objekt zu definieren, das auf dem Datensatz **cars_df** basiert und auf der X-Achse **KM** darstellt und **Price** auf der Y-Achse.
- ☐ Ergänzen Sie nun eine Schicht (**geom**), die ein Streudiagramm der beiden Variablen darstellt.
- ☐ Färben Sie nun die Datenpunkte so ein, dass die unterschiedlichen Arten von Kraftstoff (**Fuel_Type**) erkennbar sind.
- ☐ Nutzen Sie **geom_smooth()** um eine gemeinsame Trendlinie (für alle Kraftstoffarten) hinzuzufügen. Sie wollen eine lineare Regressionsfunktion haben (nicht den Standard). Nutzen Sie die Hilfefunktion **?geom_smooth()** um herauszufinden, wie Sie diese erhalten können.
- ☐ Nutzen Sie nun das **shape**-Attribut, um die Kraftstoffarten auch in der Form unterscheidbar zu machen.

- ☐ Nutzen Sie **labs( )**, um Ihrer Abbildung einen Titel, einen Untertitel sowie sinnvolle Achsenbeschriftungen zu geben. Benennen Sie auch den Titel der Datenlegende korrekt.
- ☐ Nutzen Sie die geoms: **geom_hline()** und **geom_vline()** um eine horizontale und eine vertikale Linie bei den oben berechneten Mittelwerten einzuzeichnen. Färben Sie die Linien **grau** ein und wählen Sie eine **gepunktete Linie**.
- ☐ Ändern Sie die Farbpalette in der Abbildung auf die Farbpalette mit den Farben aus Ihrem Corporate Design und formatieren Sie Ihre Abbildung mit dem Theme: **theme_bw()**.
- ☐ Erzeugen Sie eine zweite Abbildung, in der der Zusammenhang zwischen Price und KM nach Kraftstoffarten getrennt angezeigt wird.
- ☐ Speichern Sie das R-Skript ab und prüfen Sie, ob Ihre Ergebnisse replizierbar sind.