Hausaufgabe III: Datentransformation und Modellierung

1. Vorbereitendes

a. Erstellen Sie ein neues Verzeichnis mit dem Namen `GDSR_3`, definieren Sie dieses Verzeichnis als Ihr Arbeitsverzeichnis und speichern dort ein neues R-Skript mit dem Namen `Hausaufgabe_3.R`.

b. Ergänzen Sie in Ihrem R-Skript folgende erste Zeile: `rm(list=ls())`

c. Installieren Sie das Paket `nycflights13` aus dem wir die Daten nutzen werden.

d. Laden Sie die Pakete `tidyverse` und `nycflights13`.

e. Betrachten Sie den über `nycflights13` geladenen Datensatz `flights`.

f. Informieren Sie sich über den Datensatz mit dem Befehl `?flights`. Der Datensatz enthält Informationen zur Pünktlichkeit von allen Flügen, die 2013 von einem der drei New York City-Flughäfen gestartet sind.

g. Der Datensatz `flights` enthält knapp 10.000 Beobachtungen, in denen eine oder mehrere Informationen fehlen. Erzeugen Sie ein neues R-Objekt mit dem Namen `Fluege_df`, in dem Sie alle vollständigen Beobachtungen des Datensatzes `flights` ablegen. Arbeiten Sie im Folgenden mit dem `Fluege_df` Datensatz weiter.

2. Analyse

Nutzen Sie für alle Aufgaben den Pipe-Operator soweit wie möglich:

a. Erstellen Sie eine Tabelle mit der Anzahl an Flügen je NYC-Flughafen (`origin`), die eine Flugzeit (`air_time`) von über 120 Minuten haben.

b. Erstellen Sie eine Tabelle mit den folgenden Informationen je NYC-Flughafen: Durchschnittliche Flugzeit, minimale Flugzeit, maximale Flugzeit, Summe der Minuten in der Luft, Anzahl der Flüge.

c. Ermitteln Sie für jeden Flughafen die Verbindung mit der längsten Flugzeit und das zugehörige Ziel (`dest`).

d. Erstellen Sie eine Darstellung die drei Boxplots (getrennt nach Flughäfen) enthält, die die Verteilung der Variable (`air_time`) darstellen. Achten Sie auf eine korrekte Beschriftung der Achsen und vergeben Sie einen Titel für die Abbildung.
e. Erstellen Sie die gleichen Abbildungen noch einmal, aber jetzt nur für Flüge mit mehr als 60 min Verspätung bei Abflug (`dep_delay`). Achten Sie auf eine korrekte Beschriftung der Achsen und vergeben Sie einen Titel für die Abbildung.
f. Schätzen Sie ein lineares Regressionsmodell mit der abhängigen Variable `air_time` und den erklärenden Variablen `distance` und `origin` und lassen Sie sich den detaillierten Output anzeigen.
g. Schätzen Sie das Modell aus Teilaufgabe f) noch einmal getrennt für die drei Airlines (`carrier`) mit den meisten Flügen im Datensatz.
h. Speichern Sie das R-Skript ab und prüfen Sie, ob Ihre Ergebnisse replizierbar sind.