Wintersemester 2020/21
Christoph Herrmann

Aufgabe 1:
Laden Sie das Paket *tidyverse* in Ihren Workspace und betrachten Sie erneut den Datensatz *starwars*. Im Folgenden sollen Sie lernen, mit dem Paket *stringr* umzugehen.

(a) Der Datensatz *starwars* enthält mehrere Variablen des Datentyps *character*. Verschaffen Sie sich einen Überblick über die Merkmale *hair_color*, *eye_color* und *skin_color*, in dem Sie sich die durchschnittliche Zeichenlänge und deren Standardabweichung ausgeben lassen. Sind die durchschnittlichen Zeichenlängen und Standardabweichungen der drei Variablen unterschiedlich?

(b) Schreiben Sie vor die Ausprägung der Variable *hair_color* den Präfix *hair_color:* und verbinden Sie die Zeilen 1 bis 10 und 51 bis 60 zu einem Vektor. Das Ergebnis eines Elementes des Vektors müsste wie folgt aussehen: *hair_color: blond*.

(c) Verbinden Sie die Spalten *hair_color* und *eye_color* miteinander und verwenden Sie zur Verbindung der Ausprägungen ein *and*.

(d) Erzeugen Sie eine neue Spalte im Datensatz *starwars* und benennen Sie diese *short_hair*. Die Variable *short_hair* soll die ersten zwei Buchstaben des *characters hair_color* enthalten.

(e) Erzeugen Sie einen neuen Datensatz *star_string*, der die Spalten *name, hair_color, eye_color, skin_color, homeworld, films, vehicles* und *starships* des Datensatzes *starwars* enthält. Erzeugen Sie außerdem eine Spalte, die Sie *films_low* nennen, in der alle Buchstaben der Spalte *films* kleingeschrieben sind.

(f) Sortieren Sie die *strings* der Variable *films_low* nach dem deutschen Alphabet. Was stellen Sie fest?

(g) Beheben Sie das Problem, in dem Sie alle Zeichen entfernen (außer Leerzeichen), die nicht zu den Filmnamen gehören und überschreiben Sie die Spalte. Sortieren Sie die Spalte erneut.

(h) Lassen Sie sich als nächstes alle Namen der Starwars-Charaktere ausgeben, die eine Zahl enthalten. Welche Charakter sind das?

(i) Wie groß ist der Anteil der Charaktere, die von einem Heimatplaneten kommen, der auf m, e, n oder t endet?

(j) Welche vier Starwars-Charaktere haben den “Millennium Falcon” geflogen? Nutzen Sie hierfür die Funktion *str_detect*.

Hinweis: Vergeben Sie für die jeweiligen Objekte, die Sie erzeugen, sinnvolle Objektnamen. Orientieren Sie sich hierbei an den Vorschlägen aus der ersten Lerneinheit.

Aufgabe 2:
In dieser Aufgabe sollen Sie den Umgang mit Faktoren mithilfe des Paketes *forcats* üben. Als Übungsdatensatz wird wiederum der Datensatz *starwars* aus dem Paket *dplyr* verwendet.

(a) Generieren Sie einen neuen *tibble*, in dem die Variable *hair_color* nur als Faktor enthalten ist.

(b) Sortieren Sie die Level der Variable absteigend nach ihrer absoluten Häufigkeit. Plotten Sie anschließend die Variable mithilfe des Befehls *ggplot(aes(hair_color)) + geom_bar()*. Vergleichen Sie das Ergebnis mit einem Plot, den Sie aus den ursprünglichen Daten erzeugen. Was fällt Ihnen auf?

(c) Überschreiben Sie die Spalte *hair_color* im neu erzeugten Datensatz, in dem Sie alle zuerst genannten Haarfarben zu einer Ausprägung, benannt nach der zuerst genannten Haarfarbe, zusammenfügen. Zählen Sie anschließend die absoluten Häufigkeiten der Haarfarben aus.

(d) Sortieren Sie nun die neu erzeugten Ausprägungen wieder so um, dass die Ausprägungen absteigend nach ihrer absoluten Häufigkeit sortiert sind. Erzeugen Sie anschließend ein Säulendiagramm.

(e) Nutzen Sie nun die Variable *eye_color* aus dem Datensatz *starwars*. Formen Sie die Ausprägungen so um, dass nur die sechs häufigsten Ausprägungen vorkommen, die restlichen Ausprägungen sollen unter der Faktorstufe *other* zusammengefasst sein. Zählen Sie im Anschluss die absoluten Häufigkeiten aus.

Aufgabe 3:
Nach dem Sie sich nun mit *strings* und *factors* beschäftigt haben, soll in dieser Aufgabe der Umgang mit Daten und Datumszeitangaben geübt werden. Hierfür müssen Sie das Paket *weatherData* von Github installieren. Nutzen Sie folgenden Code:

```
install.packages("devtools")
library(devtools)
install_github("Ram-N/weatherData")
library(weatherData)
newyork <- NewYork2013
```

(a) Schreiben Sie die Spalte *Time* in den Datentyp *datetime* und den Datensatz in die Datenstruktur *tibble* um.

(b) Fügen Sie nun in einem nächsten Schritt mithilfe des Paketes *dplyr* die folgenden Spalten ein:

- *date* (*Time* als Datentyp *date*)
- *year* (enthält nur das Jahr)
- *month* (enthält nur den Monat)
- *day* (enthält nur den Tag im Monat)
- *week* (enthält nur die Woche)
- *year_day* (enthält nur den Jahrestag)
- *week_day* (enthält nur den Wochentag beginnend mit Montag)
- *hour* (enthält nur die Stunde des Tages).

(c) Lassen Sie sich ein Liniendiagramm mit folgendem Code ausgeben: *ggplot(aes(x = week_day, y = Temperature, colour = factor(week))) + geom_line()*. Nutzen Sie hierfür nur die Zeilen für den Monat Januar und vor 1 Uhr morgens aus dem Datensatz.

Hinweis: Speichern Sie Ihr Skript und Ihren Workspace in einem geeigneten Ordner.