Klausur Empirisches Arbeiten

Teil Explorative Datenanalyse mit R

Prüfer	Prof. Dr. Nicolas Meseth
Semester	SS 2024
Max. Punktzahl	40 (oder $\frac{1}{3}$ der Gesamtpunktzahl)
Erlaubte Hilfsmittel	alle

Hinweise zu diesem Klausurteil

- Bitte nutzt die Datei `nachname_vorname_lösungen.R` für die Beantwortung der Fragen und fügt euren R-Code jeweils unter der Frage ein. Bitte entfernt am Ende alle Codereste, die nicht zur Antwort gehören.
- Denkt daran, eure Matrikelnummer und Namen vor der Bearbeitung in die ersten beiden Zeilen einzutragen.
- Ersetzt vor der Abgabe eure Vor- und Nachnamen im Dateinamen. Als Beispiel: `max_mustermann_lösungen.R`
- Ladet die Datei über den Abgabeordner im ILIAS-Lernraum der Veranstaltung hoch! Die Abgabe muss vor dem offiziellen Ende der Bearbeitungszeit erfolgen!

Teil 1: Datensatz "Campusbier-Bestellungen"

Im ersten von zwei Teilen könnt ihr insgesamt **20 Punkte** erreichen.

Bevor ihr mit der Bearbeitung der Aufgaben beginnt, kopiert die Datei `orders.csv` in euer Arbeitsverzeichnis und ladet den Datensatz als Tibble mit dem Namen `orders`.

```
library(tidyverse)
orders <- read_csv("data/orders.csv")
```

Aufgabe 1.1: Datentransformation

Beantwortet die folgenden Fragen mit R. Das Ergebnis soll in diesem Teil als Tabelle (Tibble) und *nicht* als Visualisierung dargestellt werden.

a) Gebt alle Spaltennamen des Datensatzes aus, die bool'sche Werte enthalten! (1 Punkt)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

b) An welchen Wochentagen macht der Campusbier-Onlineshop den meisten Umsatz? (2 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

c) Kaufen Kund*innen, die Marketing-E-Mails zugestimmt haben, im Durchschnitt mehr als andere Kund*innen? (3 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

d) Erstellt eine neue Spalte, in der ihr die Kunden anhand der Anzahl an bereits getätigten Bestellungen in 3 Gruppen einteilt: A-Kunden, die bereits 10 Mal oder häufiger bestellt haben. B-Kunden, die zwischen 4 und 9 Bestellungen getätigt haben. Und C-Kunden, die den Rest ausmachen! Zeigt im Ergebnis nur die neue Spalte und die Spalte `customer_orders_count`! (4 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

Aufgabe 1.2: Datenvisualisierung

Findet eine passende Visualisierungsform für die folgenden Fragen und erstellt diese mit R und `ggplot2`!

a) Bezahlen Männer oder Frauen relativ gesehen häufiger mit Paypal? (4 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

b) Wie entwickelt sich der wöchentliche Umsatz im Postleitzahlengebiet Haste (49090) verglichen mit dem Rest Osnabrücks in den vergangenen beiden Jahren? (6 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

Teil 2: Datensatz "YouTube"

Im zweiten Teil könnt ihr insgesamt **20 Punkte** erreichen!

Bevor ihr mit der Bearbeitung der Aufgaben beginnt, kopiert die Datei `youtube_metadata.csv` in euer Arbeitsverzeichnis und ladet den Datensatz als Tibble mit dem Namen `yt`. Die Daten beinhalten die Metainformationen zu den YouTube-Videos der sechs größten niedersächsischen Hochschulen sowie der vier größten deutschen privaten Hochschulen, gemessen an der Anzahl der eingeschriebenen Studierenden.

```
library(tidyverse)
energy <- read_csv("data/youtube_metadata.csv")
```

Aufgabe 2.1: Datentransformation

Beantwortet die folgenden Fragen mit R. Das Ergebnis soll in diesem Teil als Tabelle (Tibble) und *nicht* als Visualisierung ausgegeben werden.

a) Wie viele Videos hat jede Hochschule auf YouTube veröffentlicht? (2 Punkt)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

b) Die Videos welcher Hochschulen werden am häufigsten angesehen? Vergleicht die Hochschulen miteinander! (4 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

c) Ermittelt den prozentualen Anteil englischsprachiger Videos für jede Hochschule und gebt als Ergebnis eine absteigend sortierte Liste aus! (4 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

Aufgabe 2.2: Datenvisualisierung

a) Seit wann sind die Hochschulen auf YouTube aktiv und wie intensiv ist die Nutzung im Zeitverlauf? (5 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```

b) Wie ist die Verteilung der Längen der Videos für jede Hochschule? (5 Punkte)

```
# Fügt eure Lösung bitte in die .R-Datei unter dieser Frage ein
```