

Klausur Empirisches Arbeiten

Teil Explorative Datenanalyse mit R

Prüfer	Prof. Dr. Nicolas Meseth
Semester	WS 24/25
Max. Punktzahl	40 (oder $\frac{1}{3}$ der Gesamtpunktzahl)
Erlaubte Hilfsmittel	keine, außer 2-Seiten Spickzettel

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: `mustermann_max_lösungen.R`
- Ladet die Datei über den Abgabeordner im ILIAS-Lernraum der Veranstaltung hoch! Die Abgabe muss vor dem offiziellen Ende der Bearbeitungszeit erfolgen!
- Außer dem von euch erstellten Spickzettel sind keine weiteren Hilfsmittel erlaubt!

Teil 1: Datensatz "Tweets"

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

Bevor ihr mit der Bearbeitung der Aufgaben beginnt, ladet den Tweets-Datensatz als Tibble mit dem Namen `tweets`:

```
library(tidyverse)
tweets <- readRDS("data/tweets/tweets_ampel.rds")
```

Aufgabe 1.1: Datentransformation

Beantwortet die folgenden Fragen mit R und dem Tidyverse. 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! (2 Punkte)

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

b) Listet alle *originären* Tweets von Olaf Scholz auf, die mehr als 10.000 Likes bekommen haben! (4 Punkte)

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

c) Vergleicht die durchschnittlichen Likes aller Tweets mit dem Hashtag #Ukraine mit denen ohne diesen Hashtag. Verwendet eine Kenngröße, die gegen Ausreißer robust ist! (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) Wie ist die Verteilung der Likes über alle Tweets hinweg? Entfernt vor der Visualisierung oberen 10% der Tweets mit den höchsten Werten für `favorite_count`, unter der Annahmen, dass dies Ausreißer sind! (5 Punkte)

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

b) Wie viele Tweets versendet Christian Lindner durchschnittlich in jeder Stunde eines Tages? (5 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, ladet die YouTube-Daten in einen Tibble mit dem Namen `yt`:

```
library(tidyverse)
yt <- read_csv("data/youtube/youtube_unis.csv")
```

Der Datensatz beinhaltet Metadaten zu YouTube-Videos der größten niedersächsischen Hochschulen sowie der vier größten privaten Hochschulen deutschlandweit. Die Daten wurden am 29.11.2024 abgerufen.

Folgende Variablen sind im Datensatz enthalten:

Variablenname	Beschreibung
<code>yt_id</code>	Eindeutige ID eines Videos
<code>channel_id</code>	Eindeutige ID des YouTube-Channels
<code>channel_title</code>	Name des YouTube-Channels
<code>published_at</code>	Zeitpunkt der Videoveröffentlichung
<code>url</code>	Link zum Video auf YouTube
<code>title</code>	Titel des Videos
<code>description</code>	Kurze Beschreibung des Videos
<code>duration_seconds</code>	Dauer des Videos in Sekunden
<code>views</code>	Anzahl der Ansichten zum Zeitpunkt des Datenabrufs
<code>language</code>	Die Sprache des Videos
<code>thumbnail</code>	URL zum Vorschaubild des Videos
<code>likes</code>	Anzahl der Likes des Videos zum Zeitpunkt des Datenabrufs
<code>comment_count</code>	Anzahl der Kommentare zum Zeitpunkt des Datenabrufs

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) **Welche Hochschulen sind im Datensatz vertreten und wie viele Videos haben sie jeweils veröffentlicht?** (2 Punkte)

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

b) Wie lang sind die Videos jeder Hochschule durchschnittlich in Minuten gemessen?
(4 Punkte)

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

c) Berechnet für jede Hochschule die neue Kennzahl *Likes per View* und stellt den Durchschnittswert für jede Hochschule tabellarisch dar (4 Punkte)

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

Aufgabe 2.2: Datenvisualisierung

a) Wie entwickelt sich die durchschnittliche Anzahl Likes für jede Hochschule über die letzten Jahre ab 2015 inklusive? Erstellt eine Visualisierung, die einen schnellen Vergleich zulässt. (5 Punkte)

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

b) Gibt es einen Zusammenhang zwischen der Anzahl Views und den Likes, die ein Video erhält? (5 Punkte)

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

Viel Erfolg!