Klausur Empirisches Arbeiten

Teil Explorative Datenanalyse mit R

Prüfer	Prof. Dr. Nicolas Meseth
Semester	WS 22/23
Max. Punktzahl	40 (oder $\frac{1}{3}$ der Gesamtpunktzahl)
Erlaubte Hilfsmittel	alle

Hinweise zu diesem Klausurteil

- Bitte nutzt das die Datei `nachname_vorname_lösungen.R` für die Beantwortung der Fragen und fügt euren R-Code jeweils unter die 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 "Klausurteil Meseth" 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("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, denen wir Marketing-Mails schicken dürfen, 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 "KI-Modelle"

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

Bevor ihr mit der Bearbeitung der Aufgaben beginnt, kopiert die Datei `ai_data.rds` in euer Arbeitsverzeichnis und ladet den Datensatz als Tibble mit dem Namen `ai`. Die Daten stammen aus Giattino et al. (2022).

```
library(tidyverse)
ai <- readRDS("ai_data.rds")
```

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) Erstellt einen Tibble, der nur die numerischen Spalten enthält! (1 Punkt)

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

b) Welche 5 KI-Modelle haben die meisten Datenpunkte für das Training verwendet? (2 Punkte)

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

c) Erstellt eine Übersicht der Spalten sortiert nach ihrem Füllgrad. Die am schlechtesten gefüllten Spalten sollen im Ergebnis oben erscheinen! (3 Punkte)

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

d) Wie viele Sprachmodelle sind seit 2010 jeweils pro Jahr neu veröffentlicht worden? (4 Punkte)

Hinweis: Ob es sich um ein Sprachmodell handelt, könnt ihr über die Variable `domain` herausfinden.

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

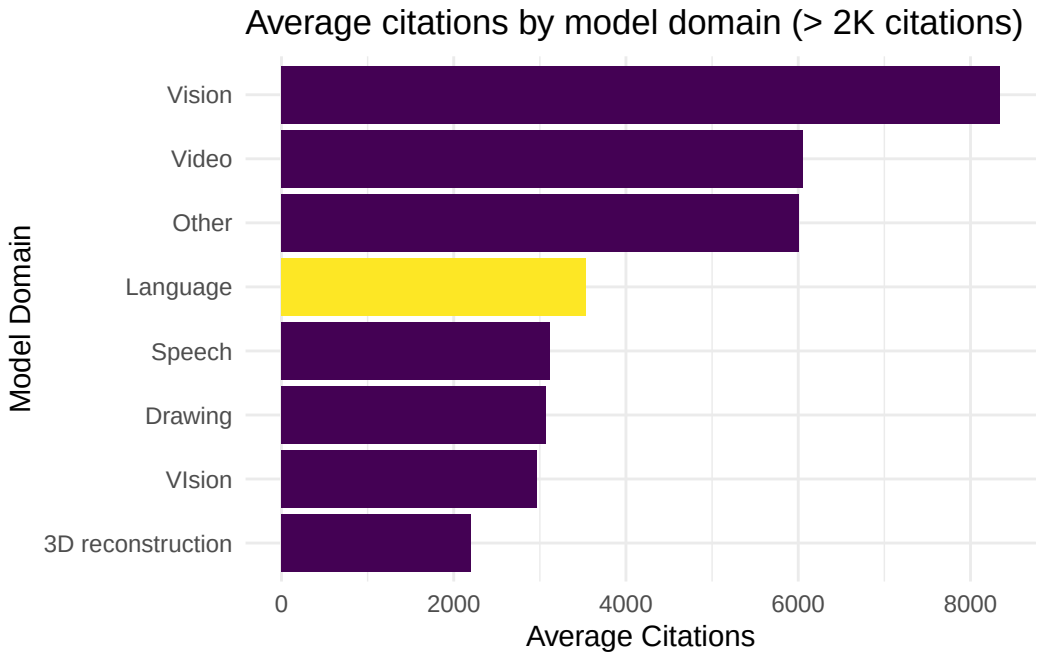
Aufgabe 2.2: Datenvisualisierung

a) Erstellt ein Punktediagramm mit den Jahren auf der x-Achse und der Anzahl Datenpunkte für das Training auf der y-Achse. Welches Problem habt ihr mit der y-Achse und wie könntet ihr es lösen? (3 Punkte)

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

b) Erstellt den R-Code für das Balkendiagramm unten und versucht dabei, das Diagramm möglichst exakt nachzubilden! (7 Punkte)

Hinweis: Die im Diagramm verwendete Farbpalette ist `viridis` für diskrete Skalen.



Quellen

Giattino, Charlie, Edouard Mathieu, Julia Broden, and Max Roser. 2022. "Artificial Intelligence." *Our World in Data.*